Bioinformatics Analysis and Annotation of Microtubule Binding and Associated Proteins (MAPs) – Creating a Database of MAPs

A Thesis Submitted to the Faculty of the School of Informatics, Indiana University, Indianapolis

By

Narmada Shenoy

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

August 2005

Master's Committee

_____

Brian Guenther, PhD

_____

Narayanan Perumal, PhD

_____

Malika Mahoui, PhD

Accepted by the Graduate Faculty, Indiana University,
in partial fulfillment of the requirements for the degree of Master's of Science in
Bioinformatics Graduate Program

To Amma, Annu, Sada and Sameer

ACKNOWLEDGMENTS

It has been a great learning experience being a graduate student at the School of Informatics, Indiana University Purdue University-Indianapolis. I would like to take this opportunity to thank the people who have helped make this thesis possible.

I would like to thank, my advisor, Dr. Brian Guenther, Department of Pathology, IU School of Medicine for his encouragement and guidance through the course of my thesis. His immense knowledge and insights provided a strong foundation for this thesis. He constantly challenged me to achieve greater heights and realize my full potential.

My special thanks to Dr. Narayanan Perumal, School of Informatics for being on my advisory committee, and for guiding me through the course of my graduate studies and the thesis. He has provided valuable inputs and has been an excellent mentor and guide.

I would like to thank Dr. Malika Mahoui, School of Informatics for being on my thesis committee and providing valuable inputs and advise on the thesis.

The painstaking effort by the members of my thesis committee to review my thesis and work is greatly appreciated.

Many thanks to the faculty, staff and colleagues at the Department of Bioinformatics, School of Informatics, for their cooperation and goodwill.

Finally, I would like to thank my parents, my husband Sadashiv and my son Sameer for their love, encouragement and support.

TABLE OF CONTENTS

Page
TABLE OF CONTENTS..... 5
FIGURES, TABLES AND FLOW-CHARTS…… 7
Abstract.....……8
I. Introduction…..... 10
I.A. Microtubules - Their Many Roles ……..10
I.A.1. Flagellae and Cilia……...11
I.A.2. Centrosomes.....…… 12
I.A.3. Vesicular Transport and Organelle Organization……13
I.A.4. Spindle Microtubule Dynamics .....……...15
I.A.5. Microtubular Activity during Mitosis .....……. 16
I.B. Tubulin the Building Blocks of Microtubules – A Brief Discussion.....……...19
I.C. Microtubule Binding and Associated Proteins.....…….. 21
I.C.1. Motor Proteins and MAPs .....……... 21
I.C.2. Kinetochore-associated Microtubule Binding Proteins……...22
I.C.3. Kinetochore Associated Non Motor Proteins……..23
I.C.4. Proteins Involved in Microtubule Stabilization……...24
I.C.5. Proteins Involved in Microtubule Destabilization……...25
I.D. Goals of this Research……...25
I.D.1. MAPs of Interest in this Research……..... 28
I.D.2. The Need to Analyze before Annotation……..……. 30
I.D.3. Analysis with a Focus on Protein Structure…….. 30
I.D.4. Protein Structure – A Brief Discussion……. 31
4.1. Protein – Secondary Structure…….……..31
4.2. Protein Motifs……... 32
4.3. Protein Domains……...… 33
4.4. Proteins – As Drug Targets…….. 33

II. Materials and Methods……..35
II.A. Annotation of the Microtubule Binding and Associated Proteins……...35
II.A.1. Brief description of the Bioinformatics Tools used……. 35
1.1. Psi-blast……..……. 35
1.2. Clustal W……...36
1.3. Protein Secondary Structure Predictor……..……37
II.A.2. Development of a Protocol for Analysis……37
2.1. A First-Pass - Initial Analysis and the Switch to the Current Pass – Modified Analysis…….…38
2.1.1. Psi-blast, First Pass…….……38
2.1.2. The Advantages and Limitations of Using Psi-blast……..40
2.1.3. Clustal W, First Pass……...40
2.1.4. Discussion of the Limitations of Using Clustal W……44
2.1.5. Secondary Structure Prediction, First Pass……47
2.1.6. Literature Analysis, First Pass……47
2.1.7. Psi-blast, Current Pass……49
2.1.8. Secondary Structure Prediction, Current Pass……49
2.1.9. Literature Analysis, Current Pass……...49

II.B. The Microtubule Binding and Associated Protein Database (MAP-DB)……..50
II.B.1. Conceptual design of the Database…….50
II.B.2. Tools Used to Create and Implement the Database……....50
2.1. Oracle and SQL Plus……..……50
2.2. Microsoft FrontPage 2003……..……51
2.3. JSP and JDBC……53
2.4. Installation and Configuration of the Apache Tomcat Web Server…….……...54
2.5. Web-site design and the Dynamics of the Web pages…….……..…55
2.6. Elucidation of Some Key Features in the Database……..…….56

III. Results……..……59
III.A. Analysis and Annotation……..…… 59
III.A.1. Analysis of Ensconsin (E-MAP-115-105)…….……...59
1.1. Psi-blast……59
1.2. Secondary Structure Prediction…….…….……...61
1.3. Analysis of Secondary Structure Prediction……..62
1.4. Clustal W……...62
1.5. Analysis of Clustal W Result……....63
1.6. Literature Analysis……..…..64
III.A.2. Analysis of Hook (homolog 3)……..... 66
2.1. First Pass…….….... 66
2.2. Current Pass……...….66
III.B. MAP-DB…….…….…….. 71

IV. Discussion……...79
IV.A.What we have achieved - Analysis and Annotation of MAPs and the MAP-DB…… 79
IV.B. Limitations Run into in this Study…… 80
IV.C. Future Directions…… 81

V. References.....……..…83
VI. Appendix
Resume……85

Figures

Page

Figure 1: Microtubule Structure……10
Figure 2: Section of the Eukaryotic Cilium or Flagellum ……..……12
Figure 3: The Centrosome…… 13
Figure 4: Organization of Cytosol and Vesicular Transport…….14
Figure 5: Role of Microtubules in Mitosis……17
Figure 6: Secondary Structure of Tubulin…….20
Figure 7: MT System in Toxoplasma……26
Figure 8: Mis-annotation of the FERM domain……29
Figure 9: Comparison of Clustal W and Psi-blast result……...44
Figure 10: MS Access Screen-shot of the database, with tables of the database (their respective fields) and the integrity constraints imposed …….……52
Figure 11: JDBC Logic…… 54
Figure 12: Format in which sequence information is displayed……...58
Figure 13: Ensconsin, Psi-blast result……61
Figure 14: Ensconsin, Secondary Structure Prediction……62
Figure 15: Ensconsin, Clustal W result……63
Figure16: Hook (homolog3), Psi-blast result…… 68
Figure17: Hook (homolog3), Clustal W result……..…….. .69
Figure 18: Screen to search for a particular protein……... 72
Figure 19: Screen that lists isoforms of the protein…… 73
Figure 20: Screen that displays details of the protein……74
Figure 21: Sequence Information……. 75
Figure 22: References and Annotation……...77
Figure 23: Domain details……..78

Flow-charts and Tables

Page

Flow-chart 1: The Initial Protocol of Analysis – First Pass……...39
Flow-chart 2: The Modified Protocol of Analysis – Current Pass……..….. 46
Table 1: Two Key References and Annotations for Ensconsin……65
Table 2: Two Key References and Annotations for Hook (homolog 3)……....70

Abstract

Microtubules have many roles in the cytoskeletal infrastructure. This infrastructure underlies vital processes of cellular life such as motility, division, morphology, and intracellular organization and transport. These different roles are carried out by the creation of different microtubule (MT) systems (such as basal bodies, centrioles, flagellum, kinetochores, and mitotic spindles). The changing roles require the cytoskeleton to be both dynamic and static in nature. Guiding these processes are a network of proteins that direct cellular behavior through their ability to bind microtubules (MTs) in a spatial- and temporal-specific manner. The identification and characterization of the suite of microtubule binding and associated proteins (MAPs) involved in MT systems is important for the understanding of the biological form and function of each MT system. This research involved the analysis and annotation of four MAPs – Ensconsin in Humans, Hook (homolog 3) in Humans, Protein Regulator of Cytokinesis 1 (PRC1) in Humans and Anaphase Spindle Elongation protein (ASE1) in yeast. A bioinformatics approach was used for the annotation and analysis. A protocol for analysis and annotation of MAPs was developed. During the process, some limitations in using bioinformatics tools and procedures were encountered. These limitations were overcome, the initial protocol was improved on and a modified protocol of analysis was developed. A database was designed and built to hold annotated information on the MAPs. We seek to disseminate this database and its functionalities as a web resource to the scientific community. It will provide an excellent forum for researchers to obtain relevant information on MT binding and associated proteins (MAPs). Infection by parasitic protozoa causes incalculable morbidity and mortality to humans and agricultural animals. In this research, we have also focused on MAPs in parasitic organisms of the Apicomplexan and Trypanosomatid genera. The protocol for analysis incorporates steps to analyze MAPs from these organisms as well. Malaria (a potentially life threatening disease) is caused by Plasmodium, an Apicomplexan parasite. This parasite is transmitted to people by the female Anopheles mosquito, which feeds on human blood. African Sleeping Sickness is an acute disease

caused by Trypanosoma brucei that typically leads to death within weeks or months if not treated. Microtubule-associated proteins (MAPs) and their alteration of the unique microtubule (MT) systems play major roles in these organisms throughout their life cycle and are required for their pathogenic mechanisms. Each parasite contains unique MT systems that will test our annotation process as well as prepare the DB for addition of other novel MT systems, such as those contained with plants. Additionally, these single cell organisms have a multistage life cycle that provide similar annotation challenges to those encountered when one considers multi-cellular organisms. Therefore, a researcher working on any MT system within the database will find useful information regardless of the organism that they are studying. This will leave us with a sub-set of MAPs from parasitic organisms in our database that are potential drug-targets.

I. Introduction

The microtubule (MT) infrastructure is responsible for many aspects of cellular life such as motility, division, chromosomal separation, morphology, polarity, structure of flagellum and cilia, intracellular organization and transport. We introduce here what microtubules are, the many roles that they play in the cell, what microtubule binding and associated proteins are and how they come into play to mold and alter MTs for their many roles.

I.A. Microtubules - Their Many Roles

Microtubules are ubiquitous cytoskeletal structures that are formed by the self-assembly of tubulin heterodimers. A microtubule is a hollow cylinder about 24 nm in diameter. Along the microtubule axis, tubulin heterodimers are joined end-to-end to form protofilaments with alternating α & β subunits. Staggered assembly of thirteen protofilaments yields a helical arrangement of tubulin heterodimers in the cylinder wall. The organization of heterodimers in the MT lattice is polarized, resulting in structural and kinetic differences at the MT ends.

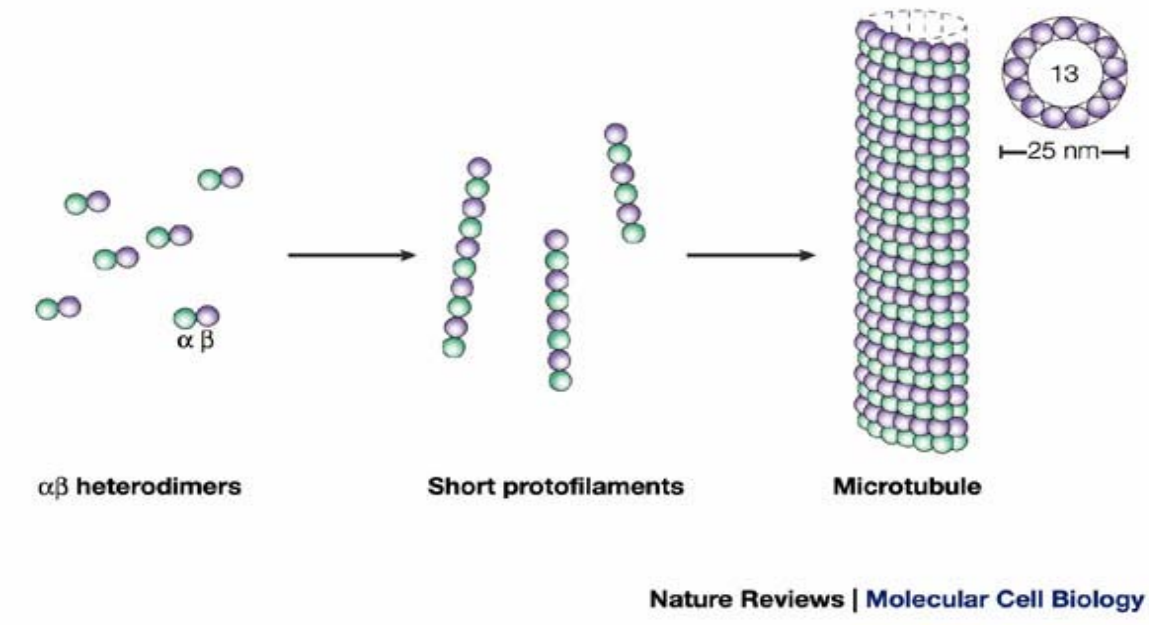


Figure 1: Microtubule Structure

The faster growing end is designated the plus end and has the beta tubulin subunit of each heterodimer exposed, whereas the slower growing end is designated the minus end and has alpha tubulin subunits exposed. In cells, the minus ends are located proximal to the centrosome or organizing center which places the plus end in the periphery of the cell. The assembly of purified tubulin *in vitro* and MT assembly *in vivo* is best described by the dynamic instability model [3]. Microtubules exhibit dynamic instability. MTs exist in persistent phases of elongation or rapid shortening, in which subunits add to or subtract from filament ends, with abrupt transitions between these two states. The switch from elongation to shortening is termed catastrophe, and the switch from shortening to elongation, rescue. Essential to microtubule functions are both their polarity and their dynamic nature. Numerous ligands also bind to tubulin affecting their assembly properties during the creation of MTs [12]. The microtubule infrastructure is responsible for many aspects of cellular life such as motility, division, chromosomal segregation during mitosis, morphology, polarity, structure of flagellum and cilia, intracellular organization and transport. Some systems in which MTs play a vital role are described here.

I.A.1. Flagellae and Cilia

Cilia and flagella are projections from the cell. They are made up of microtubules, and are covered by an extension of the plasma membrane. They are motile and designed either to move the cell itself or to move substances over or around the cell. The primary purpose of cilia in mammalian cells is to move fluid, mucous, or cells over their surface. Cilia and flagella have the same internal structure but cilia are significantly shorter in overall length. Cilia and flagella move because of the interactions of a set of microtubules inside. Collectively, these are called ‘axoneme’. A cross section of a cilium shows a circle of nine doublets, each of which has one complete and one incomplete microtubule (Figure 2). The core doublets are both complete. Extending from the doublets are sets of arms that join neighboring doublets. These are composed of the protein ‘dynein’. Nexin links are

spaced along the microtubules to hold them together. Cilia and flagella are organized from centrioles that move to the cell periphery. These are called ‘basal bodies’. Numerous cilia project from the cell membrane. Basal bodies control the direction of movement of the cilia [11].

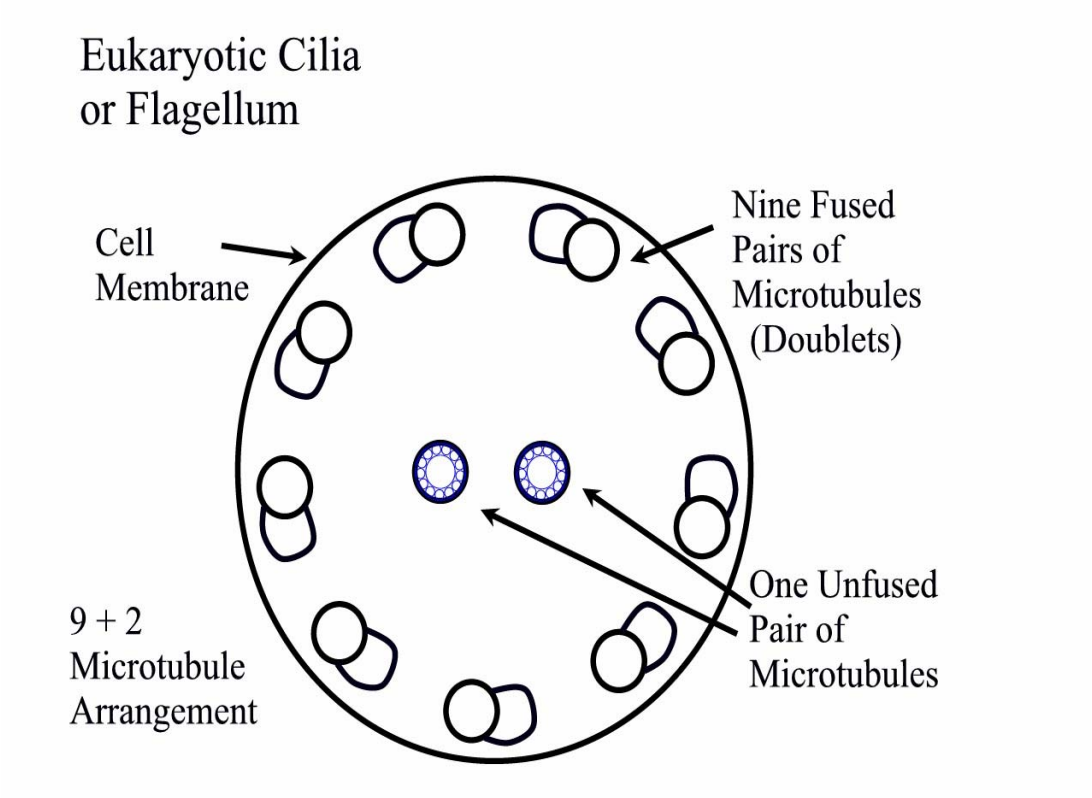


Figure 2: Section of the Eukaryotic Cilium or Flagellum

I.A.2. Centrosomes

Also known as the microtubule-organizing center (MTOC) or spindle pole, the centrosome nucleates microtubules and is important for signaling processes. Centrosomes usually contain two centrioles. Centrioles are open-ended cylinders, comprised of nine sets of triplet microtubules linked together, containing appendages on the outside and protein assemblies or sometimes vesicles on the inside. Centrosomes are the principal organizing centers for MTs in most cells. They contain many copies of the gamma tubulin ring complex (γ-TURC). Each γ-TURC is bound by one or more long, fibrous proteins to the periphery of an organelle, such as a centriole or a spindle pole body so many MTs can

grow from any one centrosome. Each γ-TURC orients the tubulin that binds to it, so the β-subunit of every polymerizing tubulin dimer lies distal to the γ-complex; consequently the plus MT end (the one that is fast to grow and shrink) is distal to the centrosome. Centrosomes, MT-associated motors, or both organize mitotic MT arrays.

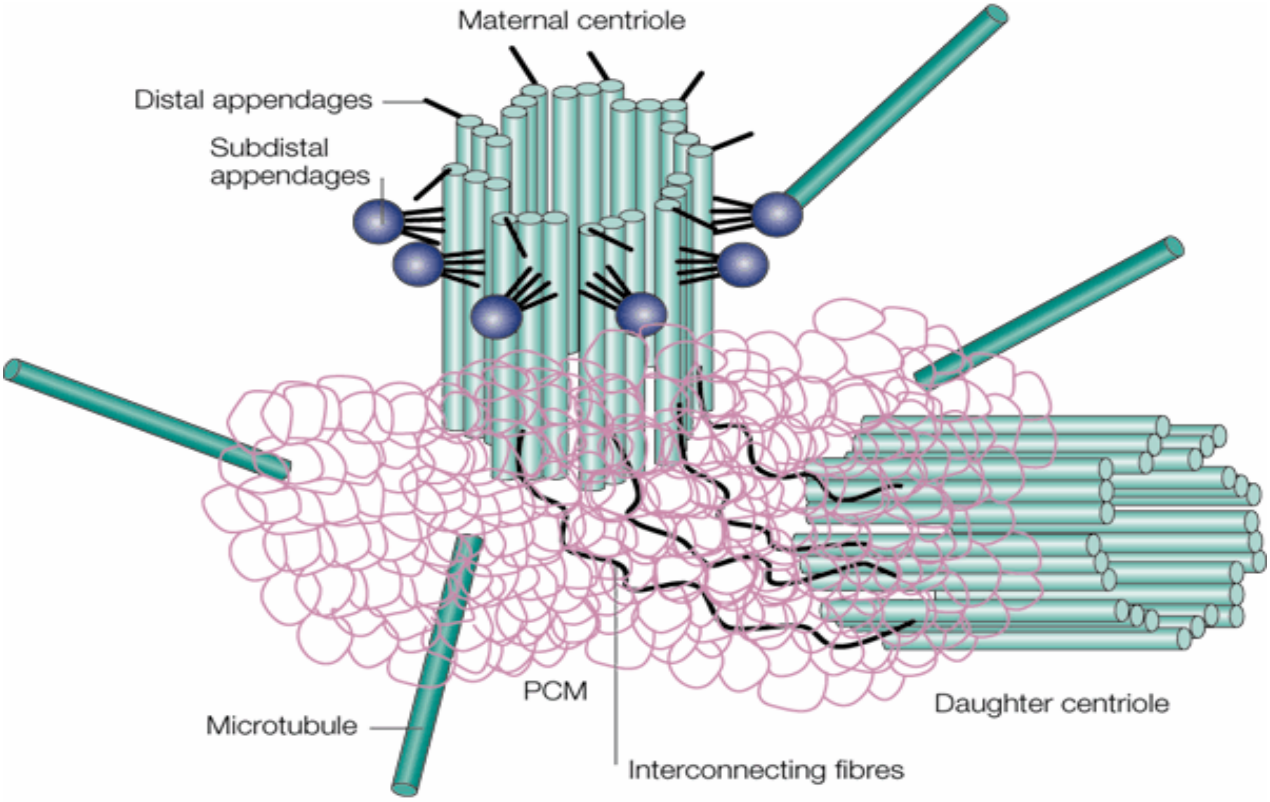


Figure 3: The Centrosome

I.A.3. Vesicular Transport and Organelle Organization

One example of the ongoing identification of MAPs and cellular function is presented by the organization of organelles. The positioning, trafficking, and architecture of the membrane compartments of the higher eukaryotic cells rely upon its MT network. Two examples of this are provided by the endoplasmic reticulum and its steady state interaction with the Golgi complex. The endoplasmic reticulum extends its tubular membrane structures towards the cell periphery along the MT scaffold. MTs are also required for the acquisition and maintenance of the Golgi complex. The

Golgi complex co localizes with the minus ends of the MTs, juxtaposed with the microtubule organizing center. Cell morphology dramatically affects the extent to which MTs are utilized for the transport of vesicles between these membranous compartments. In many cases, the distances require the use of the MTs as a highway as the rate of diffusion between the organelles is insufficient for proper functioning. This MT highway is utilized for the steady state addition and

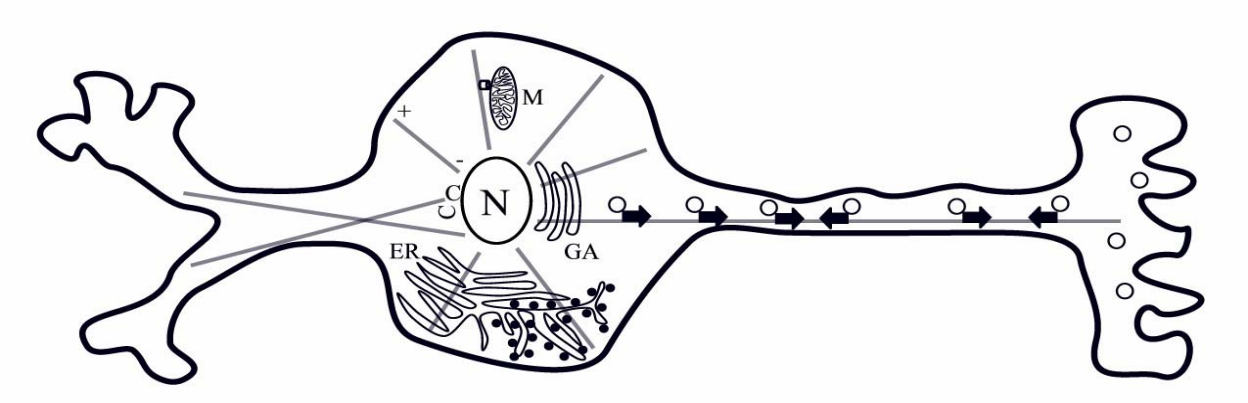


Figure 4: Organization of Cytosol and Vesicular Transport
The MTs are shown in light grey. The organelles are indicated by the following abbreviations: the endoplasmic reticulum by ER, the Golgi aparatus by GA, the nucleus by N, each centriolle by C, the mitochondria by M. The mitochondrial MT-binding anchor protein is shown as a cylinder. The motor proteins are shown as arrowheads which indicate the direction of transport. The vesicles are represented as circles.

departure of membrane vesicles between these two organelles. Correct targeting and trafficking is dependent upon the intrinsic asymmetry of the MTs which is used to generate cytosolic polarity. Two families of motor proteins mediate organelle transport in opposite directions along this MT highway. The kinesin family members carry its cargo toward the cellular periphery or plus end of the MT. The dynein family members transport membranous cargos towards the nucleus and the minus end of the MT located at the microtubule-organizing center. While these motors play an important role in the positioning of their membranous cargo, this is not the whole story. A knockout experiment of dynein showed that the mislocalized organelles were still attached to the MTs and further supported the existence of linker proteins (Harada *et al*). The identification and

characterization of the cytoplasmic linker proteins (CLIPs) between MTs and membrane vesicles and organelles has seen a flurry of activity in recent years. The number of CLIPs is much greater than the number of organelles simply due to the polarity within the organelles themselves. For example, the Golgi complex is composed of distinct stacks of cisternae (such as the cis-Golgi, cis/medial Golgi, trans-Golgi), each with a specialized role. Even within the cis-Golgi network, two CLIPs have already been identified, namely GMAP-210 and Hook3.

I.A.4. Spindle Microtubule Dynamics

In eukaryotes, morphogenesis of the microtubule cytoskeleton into a bipolar spindle is required for the faithful transmission of the genome to the two daughter cells during division [1]. This process is facilitated by the intrinsic polarity and dynamic properties of microtubules and involves many proteins that modulate microtubule organization and stability. Essential to the process of cell division is the mitotic spindle, which partitions a complete set of chromosomes to each daughter cell. The spindle consists of microtubules, polar dynamic fibers that polymerize from tubulin subunits, as well as hundreds of other proteins that function together to orchestrate chromosome segregation. These include a large set of microtubule-based motor proteins that use ATP hydrolysis to generate movement, or alter microtubule dynamics. Molecular approaches, empowered by complete genome sequences, are continuing to identify the proteins responsible for the phenomena of spindle microtubule dynamics observed.

Organizing a specific arrangement of microtubules and chromosomes within the spindle is central to how the process works [1]. Microtubules must be arranged into a bipolar array, such that each half spindle contains uniformly oriented microtubules, with their minus-ends at the pole and their plus-ends extending away. Most animal cells contain a single microtubule nucleating structure, the centrosome, having a pair of centrioles surrounded by amorphous material that harbors templates for

microtubule nucleation. The polarity of microtubule growth from centrosomes, with their minus-ends tethered and their plus-ends extending outward, facilitates proper organization of the spindle [1]. Each duplicated chromosome has a pair of specialized structures at its centromere, called kinetochores, which function to attach sister chromatids to microtubules from opposite spindle poles, to allow for directed translocation of chromosomes within the spindle. Mature kinetochores can bind and exert forces on MTs, modulate their assembly properties, and generate signals that delay anaphase onset until they have been silenced by spindle MT attachment.

I.A.5. Microtubular Activity during Mitosis

Following the creation of the mitotic spindle, the dynamic process of chromosome separation and cellular division occurs in several distinctive stages.

Prophase

By the onset of mitosis, at prophase, the centrosome and the chromosomes have duplicated and a cascade of events occurs, including nuclear envelope breakdown, chromosome condensation and centrosome separation. An increase in the frequency of microtubule shrinkage events, called catastrophes and a decrease in events rescuing growth contribute to the dismantling of the interphase array, thus allowing interaction between dynamic microtubule plus-ends and the condensed chromosomes. During prophase, chromosomes also refine the assembly of their kinetochores, and the MT-binding protein complexes that assemble on centromeric DNA [2].

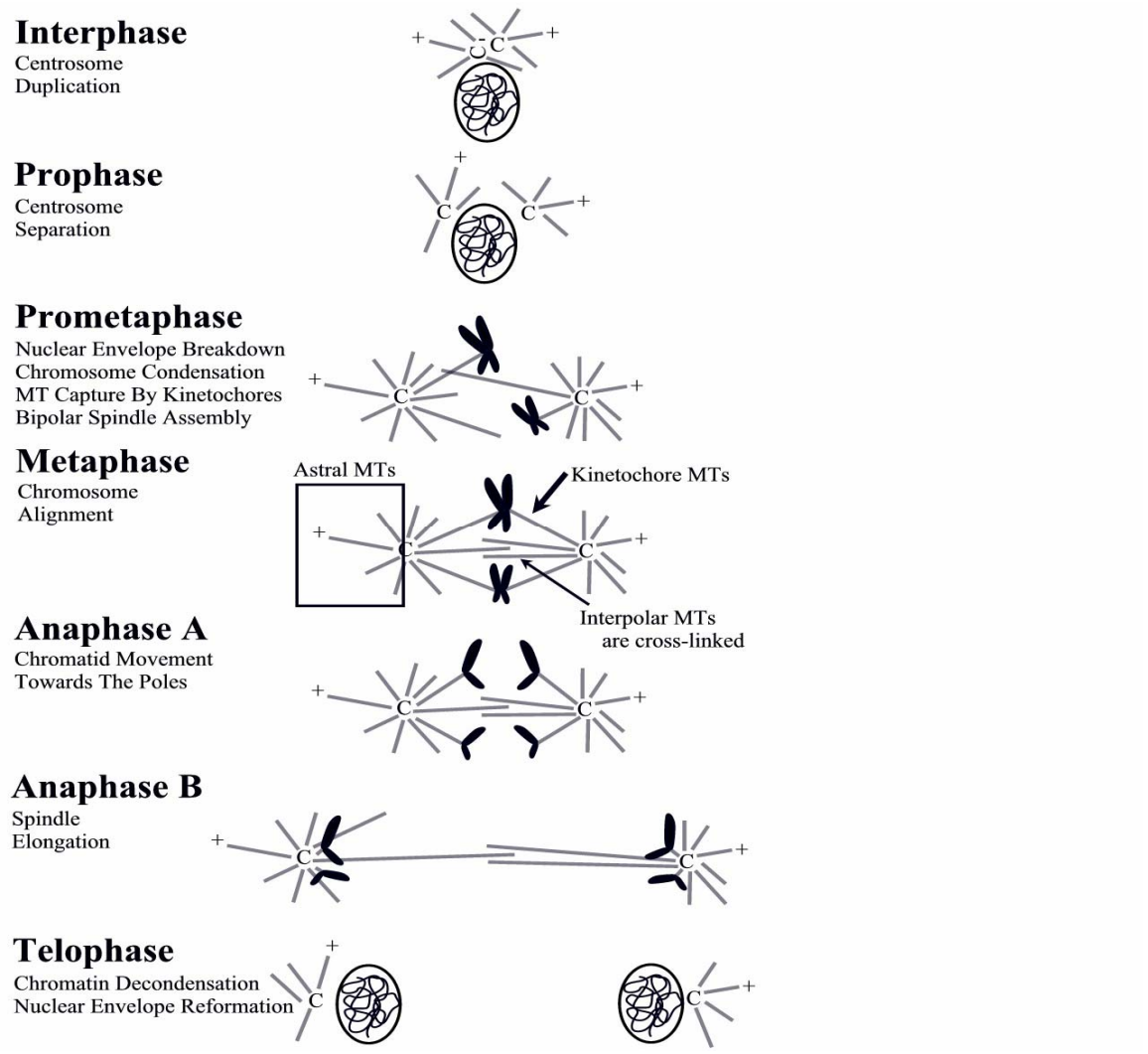


Figure 5: Role of Microtubules in Mitosis
The MTs are shown in light grey and each centriolle is indicated by the letter C.

Pro-metaphase

Pro-metaphase begins when the chromosomes start to interact with spindle MTs. Typically this occurs when the nuclear envelope breaks down. In algae and fungi, the envelope remains largely intact (closed mitosis), and prometaphase begins with the entry into and/or activation of tubulin and other spindle proteins within the nucleus. Subsequently, the polymerizing ends of many MTs come into close proximity with the chromosomes. The organization of MTs during this time is important for two reasons. First, radial arrangements enhance the likelihood that growing MTs will encounter chromosomes wherever they may lie. Second, the presence of two nearby asters, promotes the

formation of a bipolar MT array, a prerequisite for successful mitosis. Some MTs from neighboring asters interact through both motors and MT-associated proteins (MAPs) to form interpolar fibers that help to keep spindle poles apart [2]. Most pole-initiated MTs are dynamically unstable, meaning that each polymer grows and shrinks rapidly and repeatedly. During prometaphase, some microtubules emanating from one centrosome attach to the kinetochore of one of the duplicated chromatids. Subsequent attachment of the sister kinetochore to microtubules growing from the other centrosome result in the bi-orientation of the chromosome and its eventual congression to the center of the antiparallel microtubule array.

Metaphase

Once all of the chromosomes are bi-oriented and aligned, the cell is in metaphase. In addition to the kinetochore fibers, other populations of microtubules also contribute to the bipolar structure, for example, the interpolar microtubules that overlap to form an antiparallel array. The astral microtubules extend from each centrosome away from the spindle where they can interact with the cell cortex, which is the part of the microtubular lattice that lies under the cell membrane.

Anaphase

When the chromosomes are aligned and oriented, a cellular checkpoint is satisfied, and anaphase ensues as sister chromosomes separate and move toward opposite spindle poles with their kinetochores leading. Anaphase also contributes to chromosome segregation, as spindle poles separate and the central spindle forms.

Telophase

Telophase marks the reformation of the nuclear envelopes around daughter cell nuclei as the cytokinetic furrow pinches the cell into two. Spindle microtubules interact with mitotic chromosomes, binding to their kinetochores to generate forces that are important for accurate chromosome segregation.

I.B. Tubulin the Building Blocks of Microtubules – A Brief Discussion

Given the many roles of MT systems in cellular processes, it is not surprising that MTs are obligate proteinaceous elements found in all eukaryotic cells. In co-operation with other components of the cytoskeleton, namely with actin microfilaments and intermediate filaments, microtubules are involved in several basic cellular processes as segregation of genetic material, intracellular transport, maintenance of cell shape, positioning of cell organelles, extra cellular transport by means of cilia, and movement of cells by means of flagella and cilia. Though the functional impact of microtubules in cellular processes is wide-ranging, the structural unit of microtubules is to a first approximation constant; the α–β tubulin heterodimer. The atomic model of the globular tubulin proteins shows that the alpha and beta tubulin monomers have basically identical structures. Two interacting beta sheets surrounded by alpha helices form each monomer. The monomer structure is very compact, but can be divided into three functional domains. The amino-terminal domain forms a Rossmann fold (five alpha helices and six parallel beta strands), at the base of which sits the nucleotide. The intermediate domain is formed by three sequential alpha helices followed by a mixed beta sheet and two more helices, and contains the taxol-binding site. The carboxyl-terminal domain is all alpha helical and overlaps the two previous domains, making the 'crest' of the protofilament on the outside surface of the microtubule where microtubule-associated proteins and motor proteins bind. Although the last 10 residues in alpha tubulin and the last 18 residues in beta tubulin are not seen in the visualization

map, presumably because of disorder resulting from their high charge density, it is clear that they extend from a point that is near the ridge of the protofilament. These residues are the most

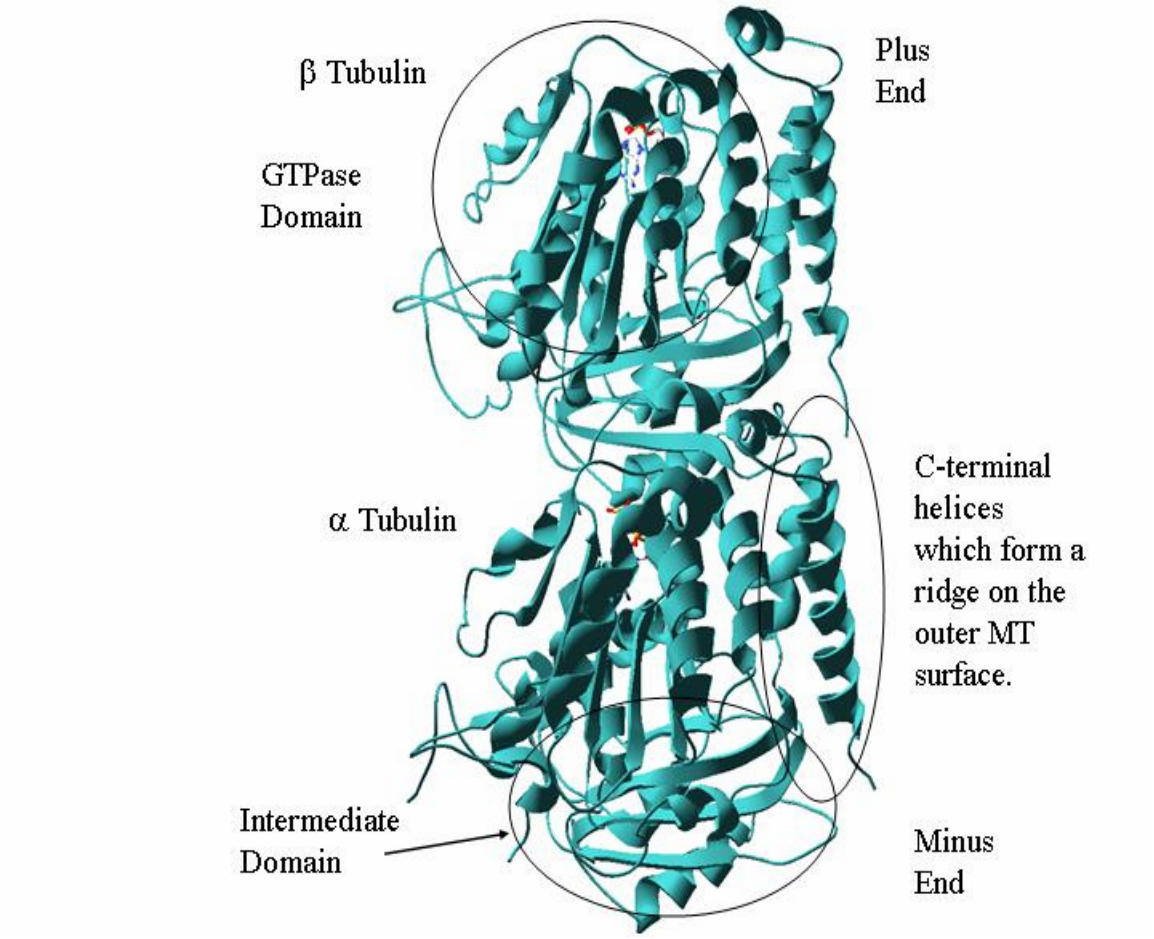


Figure 6: Secondary Structure of Tubulin

variable part of the tubulin molecule and the main determinants of isotype variety. The fact that they are exposed on the outside of the microtubule is certain to promote the idea that there are isotype-selective proteins interacting with the microtubule that may provide the functional basis for microtubule isotype variety which would be crucial to constructing specific MT systems and determining their particular function.

I.C. Microtubule Binding and Associated Proteins

The common building block (alpha and beta tubulin heterodimer) forms the basis for understanding MT-binding proteins, since this is the substrate that all these proteins must interact with and modify in order to create the MT systems responsible for motility, division, chromosomal separation, morphology, polarity, structure of flagellum and cilia, intracellular organization and transport. In order to mold and alter the MTs for all of these roles, the cell relies upon a host of MT associated proteins- a networks of proteins that direct cellular behavior through their ability to bind microtubules (MTs) in a spatial- and temporal-specific manner. Therefore, a fundamental part of understanding biological form and function requires the identification and characterization of the suite of MAPs involved in each MT system (such as basal bodies, centrioles, flagellum, kinetochores and mitotic spindles). Thus the annotation of MAPs paints a molecular portrait of each MT system. The MT binding or associated proteins commonly known as MAPs are the focus of this research.

I.C.1. Motor Proteins and MAPs

Three distinct MT processes are thought to contribute to mitotic chromosome motions: motor enzymes, the polymerization and depolymerization of MTs themselves and the flux of spindle MTs to the poles. Some motor enzymes in early mitotic movements associate specifically with kinetochores, some with chromosome arms, and some with spindle poles and the pole-initiated MTs. The initial spindle pole movements of chromosomes have been attributed to kinetochore-associated dynein, a minus end–directed motor enzyme [2]. Studies suggest that dynein plays a direct role in chromosome attachment to the spindle and to subsequent chromosome movement. Some evidence implicates dynein in the chromosome-to-pole movements of anaphase. Motors may also contribute to the away from pole motions that follow the initial kinetochore attachment to spindle MTs. Kinetochores in many organisms are associated with a plus end–directed motor enzyme: e.g., CENP-E in vertebrates. Other common kinetochore-associated motor enzymes are members of the Kin I

family, e.g., MCAK in mammals and XKCM1 in *Xenopus*). These motors depolymerize MTs by binding to and destabilizing their ends. The motor then detaches and recycles in an ATP-dependent manner. Motor enzymes at the spindle poles are likely to drive MT flux; in turn, flux can work on chromosomes. Several motors are concentrated at or near the poles, including dynein, homo-tetrameric kinesin like proteins (KLPs) of the BIM-C subfamily, and minus end–directed KLPs of the KAR3 subfamily, both in mammals and in yeasts. A role for BIM-C motors in flux has not been identified, but they might be tethered to the centrosome and try to walk toward the MT plus ends, reeling in the MT. Kar3p and, perhaps, members of that KLP family promote MT disassembly in vitro, with a preference for the MT minus end; thus pole-associated Kar3p-like motors may shorten MTs and contribute to flux. Antibody injection experiments suggest that motors of this family cooperate with NuMA to form a mechanical connection between MTs and the centrosome [2]

I.C.2. Kinetochore-associated Microtubule Binding Proteins

Studies on kinetochore-associated microtubule –binding proteins that affect tubulin dynamics, show that three classes of observation are pertinent: (*a*) CENP-E antibodies can affect MT depolymerization-dependent chromosome motion in vitro; *b*) (kinesin-like motor proteins) KLPs can couple microspheres to disassembling MTs, producing movement in the absence of soluble nucleotide, and some sphere-bound KLPs increase the rate of MT disassembly as much as fivefold; (*c*) KLPs from several subfamilies promote MT disassembly, in vivo, in vitro, or both [2]. These KLP motors appear to catalyze changes in tubulin polymerization at MT ends, and therefore can be categorized as "exotubulases." In this role, they presumably use energy from soluble nucleotides to drive MT dynamics in ways that are not governed simply by tubulin concentration. Conversely, motors that remain attached to depolymerizing MT ends can help to transduce the energy stored in the MT lattice into work. There are numerous protein complexes and enzyme activities that localize to kinetochores and probably help to regulate kinetochore-microtubule behavior. The interactions

between MAPs and motors suggest that distinctions between the factors that move on MTs and those that modulate MT dynamics are not simple. The surmise is that motors and MAPs combine to form the functionally significant attachments for MTs to kinetochores and poles [2].

I.C.3. Kinetochore Associated Non Motor Proteins

The kinetochore association of some nonmotor proteins that modulate MT dynamics is suggestive that these proteins may also contribute to the dynamic attachments between chromosomes and spindle fibers. XMAP215 is a spindle-associated MAP that promotes plus end MT growth in *Xenopus* egg extracts, humans, and fruit flies (reviewed in Ohkura et al. 2001). Homologs of XMAP215 are important for mitosis in budding and fission yeasts, where they localize at or near kinetochores and contribute to chromosome segregation. Another kinetochore-associated MAP, EB1, was first identified through its binding to the product of the tumor suppressor gene, *Adenometous Polyposis Coli* (*APC*); it is responsible for linking this complex to MT plus ends (Su et al. 1995). In mammalian cells, EB1 and APC accumulate at the kinetochore, where they interact with both motors and proteins that regulate mitotic progression: APC is a high-affinity substrate for the kinetochore-associated kinase, Bub1p, and EB1 co-immunoprecipitates with the dynein/dynactin complex (Kaplan et al. 1995). Homologues of EB1, but not APC, have been discovered in fission (Beinhauer et al. 1997) and budding yeasts (Schwartz et al. 1997), where they bind MTs and are important for high-fidelity chromosome transmission, suggesting that EB1 and its associates are important for mitosis in many organisms. CLIP170 and its homolog also associate with MT plus ends (Perez et al. 1999) and play some role in kinetochore behavior in vivo (Dujardin et al. 1998, Lin et al. 2001). Dam1p is a MT-binding protein in budding yeast; it forms a complex with four kinetochore-associated binding partners and immunoprecipitates with centromeric chromatin. This complex is important for spindle formation, and it interacts with the protein kinases Mps1p (Jones et

al. 1999) and Aurora (plus its binding partner, INCEMP) (Kang et al. 2001). The yeast Aurora kinase helps to control both MT-kinetochore associations and aspects of kinetochore maturation [2].

I.C.4. Proteins Involved in Microtubule Stabilization

Microtubule stabilizers include GTP, Mg2+ or microtubule-associated proteins (MAPs), destabilizers include GDP or elevated ionic strength. K+ at intracellular concentrations noticeably increases the stability of tubulin-MAP oligomers, in contrast to Na+. ATP and the non-hydrolyzable analogue AMP-PNP enhance oscillations by mechanisms that are not directly linked to the role of nucleotide hydrolysis in assembly. A well characterized MT-associated protein (MAP) in non-neuronal mammalian cells is MAP4. This MAP stabilizes MTs by promoting frequent rescues**.** MAP4 binding to MTs may be regulated by mapmodulin; a soluble protein which binds MAP4 and prevents MT-binding. Other proteins with MT stabilizing functions have been isolated from diverse organisms, including XMAP230 and XMAP310, from *Xenopus*. XMAP230 is a potent MT stabilizer; it increases MT elongation rate and greatly suppresses the rate of catastrophes in vitro [3].The protein is absent from prophase spindles, becomes localized to mitotic spindles at metaphase, but does not associate with astral MTs until late anaphase/telophase (Andersen and Karsenti 1997). XMAP310 also shows a striking redistribution during the cell cycle. This rescue-promoting MAP is localized to the nucleus during interphase and associates with MTs only after nuclear envelope breakdown and entry into mitosis. A third *Xenopus* MAP, XMAP215, promotes fast MT elongation *in vitro* and generates the assembly of long, dynamic microtubules (Vasquez et al. 1994). However it does not reduce catastrophes or stimulate rescues. Several proteins related to XMAP215 have been identified like TOGp from humans, ZYG-9 from Caenorhabditis elegans, Stu2p from Saccharomyces cerevisiae, and p93dis1 from Schizosaccharomyces pombe. Most of these homologs show cell-cycle-dependent MT localizations. For example, TOGp localizes to the endoplasmic reticulum during interphase but to spindle poles and MTs during mitosis (Charrasse S et al. 1994). Similarly,

ZYG-9 is cytosolic during interphase and is then localized to spindle poles throughout mitosis, and to spindle MTs until early anaphase (Matthews L R et al. 1998).

I.C.5. Proteins Involved in Microtubule Destabilization

Several proteins have been identified which can destabilize MTs. These include two proteins which may be responsible for frequent plus-end catastrophes *in vivo*, XKCM1 and Op18 (Belmont L et al. 1996). Kar3p, can destabilize MT minus ends (Endow S 1994). XKCM1 is a microtubule destabilizing kinesin that increases MT catastrophes approximately fourfold in *Xenopus* extracts; using an ATP-dependent mechanism (Walczak CE et al. 1996). Over expression of a related human protein, MCAK, results in some loss of MT polymer; this is observed during both mitosis and interphase (Maney T et al. 1998). XKCM1 and MCAK are soluble during interphase and a fraction of these proteins also localizes to spindle poles and the centromere regions of chromosomes (Walczak CE et al. 1996). Op18 was also identified as a MT destabilizer in *Xenopus* egg extracts. It is a soluble protein and binds tubulin dimers, primarily through an interaction with alpha tubulin. Op18 acts by affecting the GTP state of the microtubule [21]**.** Purified Op18 was capable of inducing microtubule catastrophes at the plus ends of GTP microtubules [21].

I.D. Goals of this Research

MAPs play a fundamental role throughout biological systems. The specificity of the MT-regulatory network relies on the unique characteristics of different MT-binding surfaces embodied by a number of critical proteins. Characterizing these surfaces is a key first step in providing the information necessary for molecular intervention. The focus here is on utilizing a bioinformatics approach to identify the MT-binding and MT-associated proteins involved in creating and maintaining the unique MT structures, on characterizing how these proteins alter and stabilize the MT structures and also on

identification of MT-binding proteins that are unique to a specific MT structure and provide the opportunity for selective molecular intervention. Single cell organisms of the Apicomplexa and

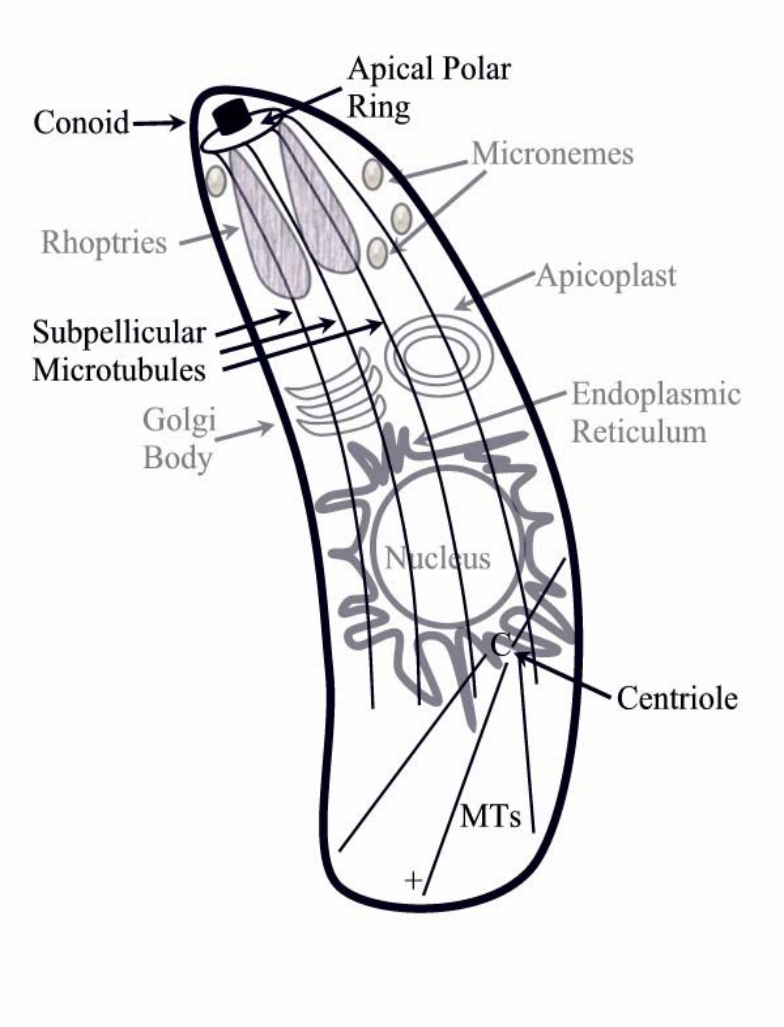


Figure 7: MT System in Toxoplasma
The cellular structures are shown in light grey and the MT systems are shown in black. Comparision to Figure 4 highlights some of the unique MT structures in Apicomplexans such as the conoid, the apical polar ring, and the subpellicular MTs.

Trypanosoma genera have unique microtubule systems, with their corresponding unique molecular portraits of MAPs to create each MT structure that are required for their life cycle. These unique MT based structures provide many opportunities to test and expand our annotation methods.

Our objectives are: 1) To annotate MT binding and MT associated proteins from all eukaryotes. Literature as well as web-based resources will be scanned for new MT-associations. Primary sequence analysis will determine whether the protein is added to one of the existing MAP families or

represents a new family of MT-binding proteins. 2) To design, develop, build, and maintain a database (DB) of MAPs based on the utility requirements of the users. This DB aims to eventually include and house MAPs within the Apicomplexan (Plasmodium, Toxoplasma, *etc.*) and Trypanosomatid genera. Plasmodium, an Apicomplexan parasite causes the disease Malaria. Toxoplasma gondii, a protozoan parasite causes toxoplasmosis. Immunocompromised individuals are very susceptible. Infection by parasitic protozoa causes incalculable morbidity and mortality to humans and agricultural animals. Microtubule-associated proteins (MAPs) and their alteration of the unique microtubule (MT) systems play major roles in these organisms throughout their life cycle and are required for their pathogenic mechanisms. The DB will also allow us to screen out parasite MT-binding proteins by demarcating the ones that they share in common with humans. This will leave us with a subset of parasite specific MT-binding proteins, which could be potential specific drug targets. Another objective is to perform a primary sequence search of these genomes with each unique MAP family defined by annotation, in order to provide additional annotation and constantly evolve as more annotated data becomes available. 3) To disseminate this DB and the related functionalities as a web resource for the scientific community. The developed application strives to provide an excellent forum for researchers to obtain relevant information on MT binding and associated proteins. In this web dissemination service, data from other species as well as that from Apicomplexan and Trypanosomatid parasites will also be included using schema and data descriptions. Genomic and proteomic information have led to a rapid increase in the identification and biochemical characterization of MAPs but there is no central database correlating this information within each MT system. This is a first effort in creating a DB on MT systems. Moreover, the single cell organisms of interest in the Apicomlexan and Trypanosomatid genera have a multistage life cycle that provides similar annotation challenges to those encountered when one considers multi-cellular organisms. This will have broad application due to the shared nature of MT

based systems throughout all organisms. In fact, an objective of this research is focused upon annotating existing MAP information from other organisms as well.

I.D.1. MAPs of Interest in this Research

The laboratory of Dr. Brian Guenther, Department of Pathology and Laboratory Medicine, Indiana University, has collected a set of 110 putative unique MAPs, based upon primary sequence analysis, and from this set, the following four MAPs were picked for annotation and analysis:

Ensconsin (E-MAP-115-105 in Humans): Epithelial cell proliferation and differentiation occur concomitant with striking re-modeling of the cytoskeleton. MTs play important roles in these processes, during which the MTs themselves are re-organized and stabilized by MAPs. Ensconsin (E-MAP-15-105) is a structural MAP that is expressed in epithelial cells [15].

Hook (homolog 3 in Humans): The positioning, trafficking, and architecture of various membrane compartments rely on microtubules. . For example, the microtubule organizing center (MTOC) specifies the characteristic perinuclear position of secretory and endocytic pathway components. Within each pathway, vesicular transport between compartments is organized along microtubules: trafficking from early to late endosomes depends on microtubules, as does retrograde and anterograde trafficking between the Golgi complex and the ER. Hook proteins constitute a novel family of microtubule-binding proteins that may link membrane compartments to microtubules. Hook (homolog 3 in Humans) is a member of this family of proteins [17].

Ase1 (Anaphase spindle elongation protein in Saccharomyces cerevisiae): Proper microtubule organization is essential for cellular processes such as organelle positioning during interphase and spindle formation during mitosis. Ase1 (Anaphase spindle elongation protein, Saccharomyces pombe) is a member of the conserved ASE1/PRC1/MAP65 family of microtubule bundling proteins, and functions in organizing the spindle midzone during mitosis [18].

PRC1 (Protein regulator of cytokinesis 1 in Humans): Midzone microtubules of mammalian cells play an essential role in the induction of cell cleavage, serving as a platform for proteins that play a part in cytokinesis. PRC1 (Protein regulator of cytokinesis 1 in Humans) is a mitotic spindle associated Cdk substrate that is essential to cell cleavage and is a microtubule binding and bundling protein both in vivo and in vitro [16].

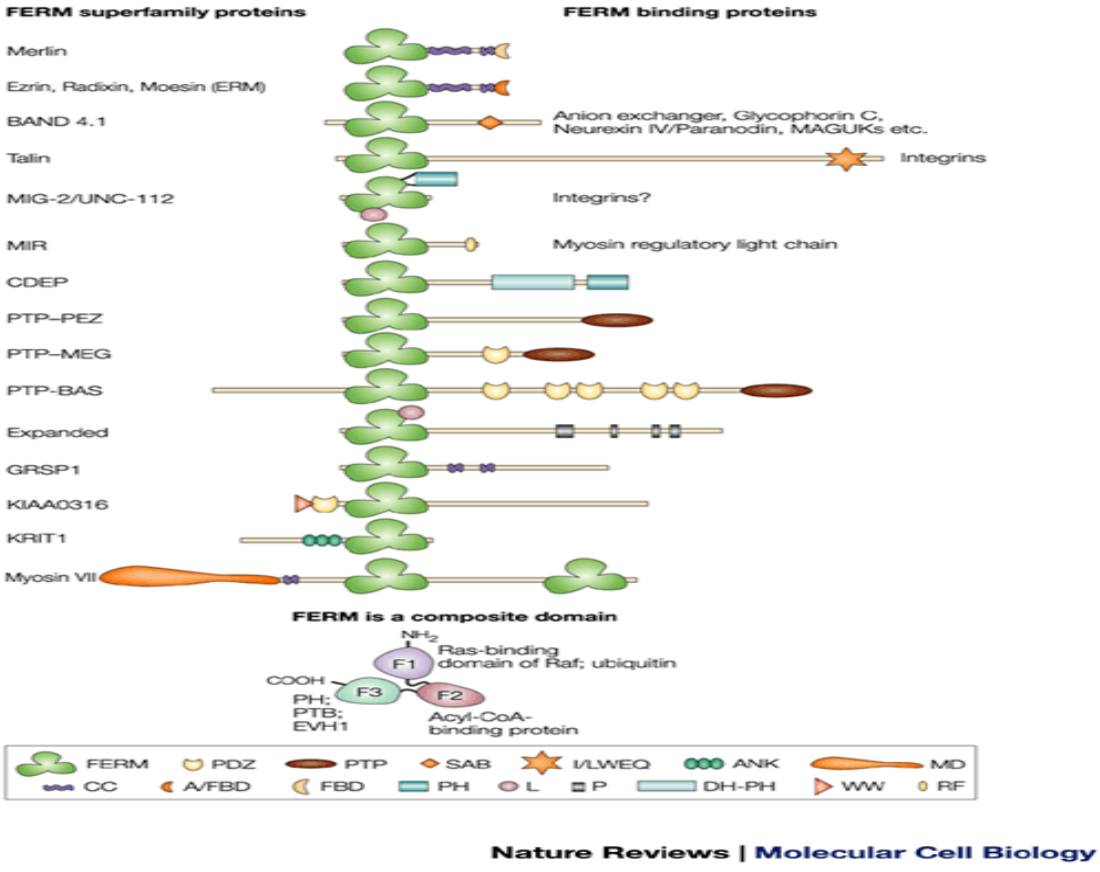


Figure 8: Mis-annotation of the FERM domain

I.D.2. The Need to Analyze before Annotation

The need to analyze the protein sequence before annotation is explained by the following example: In figure 8, the first four proteins that contain the FERM domain bind to actin. The FERM domain does not bind to actin but regulates the actin-binding function of another domain. However, the FERM domain is consistently annotated as an actin-binding domain in genbank.

I.D.3. Analysis with a Focus on Protein Structure

A challenge in this research was the application of effective bioinformatics tools to study sequence alignment and secondary structure of the MAPs to look for distant homologs that show regions of conservation across generations and to flag domain regions or conserved regions that appear to be ordered (regions with coiled –coils, helices). Limitations were encountered when computational tools such as Psi-blast, Clustal W and PredictProtein were used in the analysis. These limitations are described in detail under ‘Materials and Methods’. An initial protocol of analysis was devised and improvised upon as the analysis progressed. The aim of using a bioinformatics approach is to infer by analysis of the data obtained its biological, biomedical and evolutionary significance. We hope to encourage by the dissemination of information through our database, parallel research at the molecular level. While bioinformatics techniques give faster results, laboratory techniques though slower can be employed to validate and verify initial results and use that information to channel future research in the right direction. The essence of this analysis is to correctly identify and characterize the microtubule binding or microtubule associated domain of the MAPs and to disseminate this information without ambiguity through the MAP-DB database. A brief discussion of protein structure and the significance of its domain/domains and conserved regions is provided here to re-iterate the importance of thoroughly analyzing proteins.

I.D.4. Protein Structure – A Brief Discussion

The amino acid sequence of a protein's polypeptide chain is called its primary structure. Different regions of the sequence form local regular secondary structure, such as alpha helices or beta strands. The tertiary structure is formed by packing such structural elements into one or several compact globular units called domains. The final protein may contain several polypeptide chains arranged in a quaternary structure. By formation of such tertiary and quaternary structure amino acids far apart in the sequence are brought close together in three dimensions to form a functional region, the active site. Proteins must recognize thousands of different molecules in the cell by detailed three-dimensional interactions, which require diverse and irregular structures of the protein molecules, the most important of which is their secondary structure [5].

4.1. Protein – Secondary Structure

The main driving force for folding water-soluble globular protein molecules is to pack hydrophobic side chains into the interior of the molecule, creating a hydrophobic core and a hydrophilic surface. To bring the side chains into this core, the main chain must also fold into the interior. The main chain is highly polar and therefore hydrophilic, with one hydrogen bond donor, NH, and one hydrogen bond acceptor, C=0, for each peptide unit. These main chain polar groups must be neutralized by hydrogen bond formation in a hydrophobic environment. The formation of regular secondary structure within the interior of the protein molecule facilitates this. Such secondary structure is usually one of two types: alpha helices and beta sheets. Having the main chain NH and C0 groups participating in hydrogen bonds to each other characterize both types. All the hydrogen bonds in an alpha helix point in the same direction so the peptide units are aligned in the same orientation along a helical axis. There is a significant net dipole for the alpha helix that gives a partial positive charge at the amino end and a partial negative charge at the carboxyl end of the alpha helix. These charges attract ligands of opposite charge. Different side chains have been found to

have weak but definite preferences either for or against being in alpha helices. The most common location for an alpha helix in a protein is along the outside of the protein with one side of the helix facing the solution and the other side toward the hydrophobic interior of the protein. The second major structural element found in globular proteins is the beta sheet. This structure is built up from a combination of several regions of the polypeptide chain. Beta strands are usually from 5 to 10 residues long and are in an almost fully extended conformation and are aligned adjacent to each other such that hydrogen bonds can form between CO groups of one beta strand and NH groups on an adjacent beta strand and vice versa. The beta sheets are formed from several beta strands pleated with C alpha atoms successively a little above and below the plane of the beta sheet. The side chains follow this pattern such that within a beta strand they also point alternatively above and below the beta sheet. The secondary structural elements alpha helices and beta sheets are connected by loop regions of various lengths and irregular shape. Loop regions exposed to solvents are rich in charged and polar hydrophilic residues [5].

4.2. Protein Motifs

Simple combinations of a few secondary structure elements with a specific geometric arrangement have been found to occur frequently in protein structures. These units have been called supersecondary structures or motifs, for example the calcium binding motif in parvalbumin, calmodulin, troponin-c and other proteins that bind calcium and regulate cellular activities. Domains are built from structural motifs. The helices and beta strands of the motifs are adjacent to each other in the three dimensional structure and are connected by loop regions. Sequentially adjacent motifs or motifs that are formed from consecutive regions of the primary structure of a polypeptide chain are usually close together in the three dimensional structure. The number of such combinations found in proteins is limited and some combinations seem to be structurally favored. Thus polypeptide chains

are folded into one or several discrete units, domains which are the fundamental functional and three dimensional structural units [5].

4.3. Protein Domains

The core of a domain is built up from combinations of small motifs of secondary structures. Domains are classified into three main structural groups :1) alpha structures, where the core is built up exclusively from alpha helices 2) beta structures, which comprise antiparallel beta sheets 3) and alpha/beta structures, where combinations of beta-alpha-beta motifs form a predominantly beta sheet surrounded by alpha helices. A protein domain is a discrete portion of a protein with its own function. The combination of domains in a single protein determines its overall function [5].

4.4. Proteins – As Drug Targets

Most emphasis on structural information has been on guiding initial drug synthetic efforts. Protein structure encodes function and its sequence encodes the structure. Protein sequencing is an easier task than predicting the three dimensional structure of a protein. In the Human genome, approximately 30,000 ORFs, was determined within a couple of years. However twelve structural genomic centers determined 700 protein structures in a similar timeframe**.** A conformational change in the structure of proteins creates different surfaces for interaction with substrates and other macromolecules. Additionally, the protein molecule is dynamic. This flexibility is critical to its function and is an active area of research. The role of DNA is the preservation of information. Therefore, as a first approximation, DNA information is static and the same sequence leads to the same structure/function. Interesting questions can be addressed while drawing information from a protein sequence or structure. Like, what factors affect the information we can attach to a given sequence? On the other hand, at what level of sequence similarity can one assume that there is similar structure? Additionally, does similar structure imply similar function and sequence

conservation imply structural conservation? There is however another level of complexity because similar structure does not mean similar function, since changing a single amino acid residue can alter function. Moreover, similar function does not necessitate a similar structure. Most drugs are effective because they bind to a specific receptor site and block the physiological function of a protein. Classical drug design has been based on this concept for several decades. Compounds that are structural variants of substrates for a suitable target protein are synthesized and tested in binding studies. It is usually possible to obtain inhibitors that bind better than the physiological substrates, although several thousands of compounds may have to be synthesized and tested before a suitable drug is found. The goal is to design new drugs in a more rational way based on a knowledge of the three dimensional structure of the binding site of the relevant receptor protein. So, it is important to correctly understand which the domain/domains with in a protein are and what their respective function is. Mis-annotation must be avoided as shown in this example: in the FERM super family of proteins, four proteins that contain the FERM domain bind to actin. The FERM domain does not bind to actin but regulates the actin-binding function of another domain. However, the FERM domain is consistently annotated as an actin-binding domain.

II. Materials and Methods:

This research work involved the development of a protocol to annotate microtubule binding or associated proteins and the creation of a database to house the annotated microtubule binding and associated proteins.

II.A. Annotation of the Microtubule Binding and Associated Proteins

With the objective of populating the database with annotated MT binding and MT associated proteins, this research involved the development of a protocol to analyze and annotate four proteins known to be MT binding or MT associated. Examples of four such proteins studied here are: Ensconsin (E-MAP-115-105 in Humans), Hook (homolog 3 in Humans), PRC1 (Protein regulator of cytokinesis 1 in Humans) and Ase1 (Anaphase spindle elongation protein in Saccharomyces cerevisiae). Bioinformatics tools and techniques were used to analyze the sequence of these proteins. For analysis, the amino acid sequences of the proteins were obtained from Swiss Prot, a curated protein sequence database that is established and maintained collaboratively by the Department of Medical Biochemistry at the University of Geneva.

II.A.1. Brief description of the Bioinformatics Tools used

1.1. Psi-Blast

Position-Specific Iterated (PSI)-BLAST is a very sensitive BLAST program, making it useful for finding very distantly related proteins. More promising homologues from the many candidate proteins can be obtained by Psi-Blast. Many functionally and evolutionarily important protein similarities are recognizable only through comparison of three-dimensional structures. When such structures are not available, patterns of conservation identified from the alignment of related

sequences can aid the recognition of distant similarities. The principal design of the Position-Specific Iterated BLAST (PSI-BLAST) program provides speed, simplicity and automatic operation. The procedure Psi-blast uses can be summarized in these steps: Psi-blast takes as an input a single protein sequence and compares it to a protein database, using the gapped BLAST program. The program constructs a multiple alignment, and then a profile, from any significant local alignments found. The original query sequence serves as a template for the multiple alignment and profile, whose lengths are identical to that of the query. Different numbers of sequences can be aligned in different template positions. The profile is compared to the protein database, again seeking local alignments. Psi-blast estimates the statistical significance of the local alignments found. Finally, Psi-blast iterates an arbitrary number of times or until convergence. Profile-alignment statistics allow Psi-blast to proceed as a natural extension of BLAST; the results produced in iterative search steps are comparable to those produced from the first pass. Unlike most profile-based search methods, Psi-blast runs as one program, starting with a single protein sequence, and the intermediate steps of multiple alignment and profile construction are invisible to the user. This program is available at http://www.ncbi.nlm.nih.gov/BLAST/.

1.2. Clustal W

This is a general purpose multiple alignment program for DNA or proteins. It produces biologically meaningful multiple alignments of divergent sequences. It calculates the best match for the selected sequences and lines them up so that the identities, similarities and differences can be seen. A pairwise score is calculated for every pair of sequences that are to be aligned. These scores are presented in a table in the results. Pairwise scores are calculated as the number of identities in the best alignment divided by the number of residues compared (gap positions are excluded). The various sequences to be compared were submitted, each being separated with a '>' sign. The aligned sequences were returned to the web browser as an html file. This program can be accessed at

http://www.ebi.ac.uk/clustalw/ and can be executed using default settings or can be customized based on user needs.

1.3. Protein Secondary Structure Predictor

PredictProtein is a service for sequence analysis, and structure prediction. The PredictProtein server was developed and is maintained by the Columbia University Bioinformatics Center (CUBIC). This software takes the protein sequence as input and generates predictions of secondary structure, residue solvent accessibility, transmembrane helix location and topology, protein globularity, coiled-coil regions, cysteine bonds, and structural switching regions. Within the results obtained using this predictor, the PROFsec and COILS program results were used in our analysis. The PROFsec is a profile-based neural network program for prediction of protein secondary structure and the COILS program is a program that compares a sequence to a database of known parallel two-stranded coiled-coils and derives a similarity score. By comparing this score to the distribution of scores in globular and coiled-coil proteins, the program then calculates the probability that the sequence will adopt a coiled-coil conformation. Both these programs are local to CUBIC. When the protein sequence is submitted using a form, the results can be obtained by email. The relevant PROF sec and COILS predictions for each protein were cut and pasted to a word document and stored for analysis. The PredictProtein server is available at http://www.predictprotein.org/.

II.A.2. Development of a Protocol for Analysis

A protocol using the bioinformatics tools described, to analyze each putative microtubule binding or associated protein was devised. This protocol was tested and improvised upon to characterize, analyze and annotate the four MAPs in this research. The protocol was developed with the intention that it will be used in future research as well, to annotate other MAPs that will be added to the database eventually.

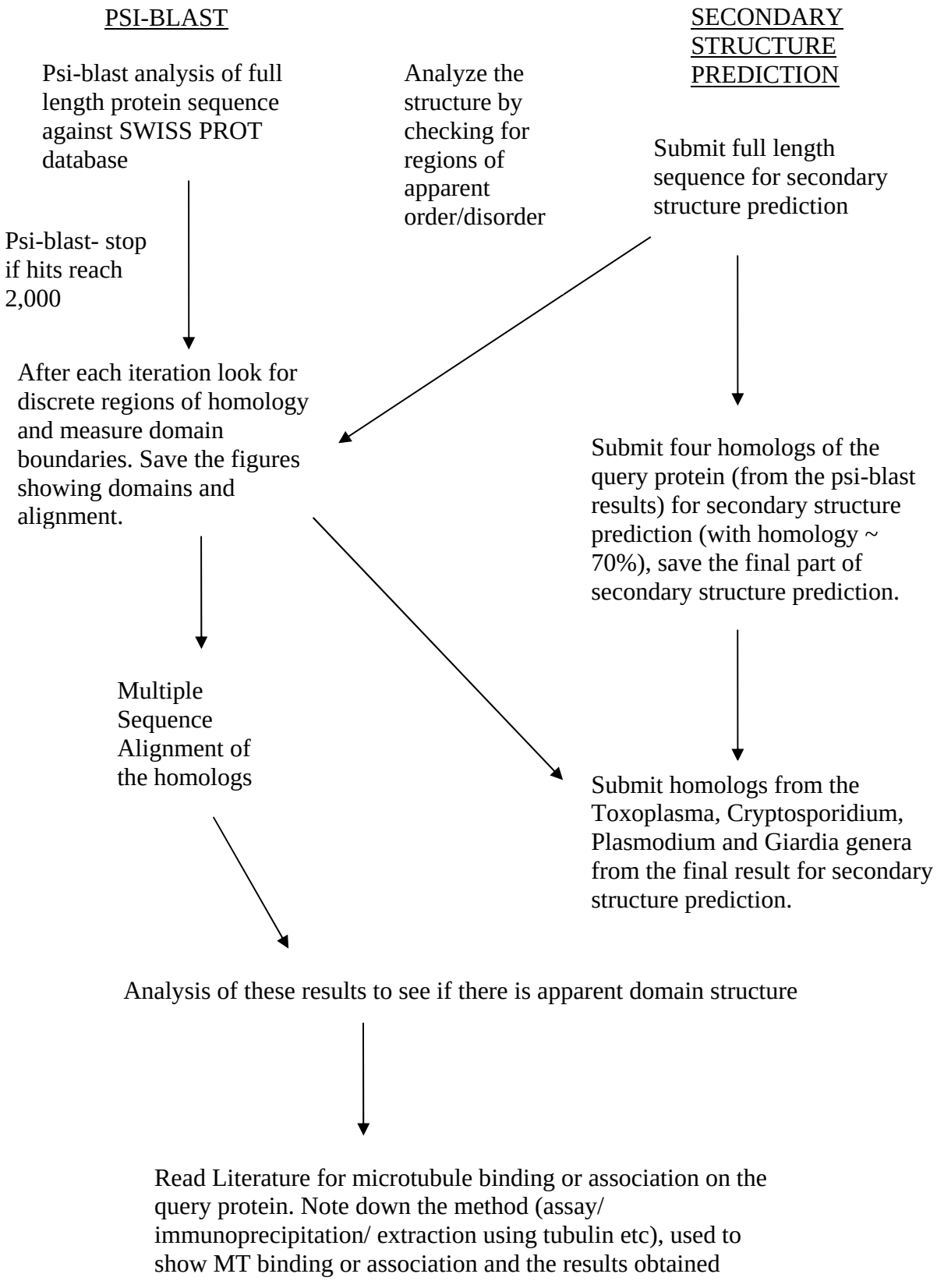
2.1. A First-Pass - Initial Analysis and the Switch to the Current Pass – Modified Analysis

The initial draft of the protocol that was designed is outlined in Flow-chart: 1. Some limitations of using bioinformatics tools in the analysis were encountered and the problems faced are described here. The protocol of analysis was improvised upon during the process and a modified draft of the protocol was devised.

2.1.1. Psi-blast, First Pass: Psi-blast analysis of the full length protein sequence against the Swiss Prot database was done. The number of hits proportionally increased after each iteration, the process became very slow and did not provide precise means for evaluating the significance of the results. To limit the number of hits and concentrate on patterns of conservation or discrete regions of homology to mark domain boundaries, it was decided to run Psi-blast iterations only until the hits reached 2000. To avoid hitting too many domains, this was later improvised upon and Psi-blast iterations were stopped if the hits were greater than 1000 (as indicated in the modified analysis). This made it easier to save result profiles and also eliminated redundant results that were obtained if the number of iterations were higher. After each iteration, the distribution of blast hits with the query sequence was scanned for regions of homology. The figures showing homologous regions and alignment were studied. Four homologs of the query protein (from the psi-blast results with homology ~ 70%), were picked for further analysis. Running psi-blast against the Swiss Prot database was also re-considered since it picked up too many false positives (non-homologous sequences that are listed as matches). It also picked up hits only for those organisms whose protein sequence entry is present in Swiss Prot and tended to miss out homologs that do not have an entry there. Therefore, Psi-blast was run against the nr (non-redundant database) in the current pass of the modified analysis.

Flow-chart:1, The Initial Protocol –First Pass is shown on the following page:

The Initial Protocol – First Pass:



2.1.2. The Advantages and Limitations of Using Psi-blast: Psi-blast provides an enormous advantage over normal blast in the detection of distantly related sequences. Iterated profile search methods have led to biologically important observations but, were quite slow and generally did not provide precise means for evaluating the significance of their results. It only works if some closely related sequences are already available, but if this is the case it finds a lot of other distantly related sequences. It provides an automated, easy-to-use version of a "profile" search, which is a sensitive way to look for sequence homologues. The Psi-blast program first performs a gapped BLAST database search. The program then uses the information from any significant alignments returned to construct a position-specific score matrix, which replaces the query sequence for the next round of database searching. Psi-blast may be iterated until no new significant alignments are found. It is used for comparing protein queries with protein databases. The results of a normal blast search are aligned and this pattern is used for the next iteration. At higher iterations, a Psi-blast profile is likely to be corrupted and false positives are identified with significant E-values. For instance in a traditional blast search one can be quite certain that a match with an E-value of 10^-13 represents a homologue; this is not clear with Psi-blast. Other test studies have indicated that profile corruptions are likely after more than five iterations. On the positive side however there are many fewer false negatives (homologous sequences that are not detected) with Psi- blast than with normal blast. False positives (non-homologous sequences that are listed as matches) are very few in normal blast but possible in Psi-blast after profile corruption. The "problem" is that the E-value reported in a Psi-blast search represents the match with the profile, not with the original sequence [19].

2.1.3. Clustal W, First Pass: Protein sequence alignments are central to the functional annotation of proteins. Four homologs of the query protein (from the psi-blast results with homology ~ 70%), were picked for further analysis. Multiple sequence alignment of these homologs was done using Clustal W. During this analysis, it was inferred that though it is expected that the alignments should

be better by using multiple sequence information, it made only small improvements to the alignment quality.

Psi-blast aligns significantly more residues correctly, whereas Clustal W did not perform as well. The limitation of the using Clustal W is shown in the following example: The output from two Clustal W results and one from Psi-blast (Figure: 9) is shown. The Clustal W alignments are of full-length protein (Protein regulator of cytokinesis 1-PRC1) from human, mouse, fruit fly, mustard weed, night-shade, bakers yeast, and brewers yeast. The two Clustal W results differ by the inclusion of two plant sequences in one and illustrate the apparent loss of conservation in the MT-binding region of the protein.

CLUSTAL W (1.82) multiple sequence alignment – 1st Result

Hs (human/homo sapien), mm (mouse/mus musculus), dminsect (fruit fly1/drosophila melanogaster), dpinsect (fruit fly 2/drosophila probo), sc (bakers yeast/Sacromyces cerevisiae), sp (brewers yeast/Sacromyces pombe)

```
test2_hs            -------------DEDAFCLSLENIATLQKLLRQLEMQKSQNEAVCEGLRTQIRELWDRL 250
test3_mm            -------------DESAFCLSLENIATLQKLLKQLEMKKSQNEAECEGLRTQIRELWDRL 250
test4_dminsect      --------------QVDHCLTPETFELLRNMQKNFADQVKELRERIDDMREKIHVLWDRL 251
test6_dpinsect      --------------DLSHNLTPETLERLRQMRNSYAEQIQELRSKIHDMREKIYVLWDRL 244
test12_sp           SHTN---------RPNDVYVTQELIDQLCKQKEVFSAEKEKRSDHLKSIQSEVSNLWNKL 277
test11_sc           LRSYGEEENSTSEIPNFHPVDRERMSKIDITLEKLQAIHKERADKKRLLMEQCQKLWTRL 420
                                      :  * :  :    .      .:       :  :   ** :*

test2_hs            QIPEEEREAVATIMSGSKAKVRKALQLEVDRLEELKMQNMKKVIEAIRVELVQYWDQCFY 310
test3_mm            QIPEEEREPVEAIMTGSKTKIRNALKLEVDRLEELKMQNIKQVIEKIRVELAQFWDQCFY 310
test4_dminsect      QETDEYAMRRVREATTYTQRTYDVLREELQRCQALRRQNLKTFIEQLRIEIKEMWDLTLK 311
test6_dpinsect      QETDESAMRRVRECTENTQRTYDILHSELQRCQALRSQNLKTFIEQLRVEISKWWDLTLK 304
test12_sp           QVSPNE-QSQFGDSSNINQENISLWETELEKLHQLKKEHLPIFLEDCRQQILQLWDSLFY 336
test11_sc           KISQEYIKTFMRNNSSLSTESLGRISKEVMRLEAMKKKLIKKLISDSWDKIQELWRTLQY 480
                    : . :         :  . .       *: : . :: : :  .:.    :: : *

test2_hs            SQEQRQAFAPFCAEDYT--------ESLLQLHDAEIVRLKNYYEVHKELFEGVQKWEETW 362
test3_mm            SQEQRQAFAPYYSEDYT--------ENLLHLHDAEIVRLRNYYDAHKELFQGVQKWEESW 362
test4_dminsect      SQQERKRFSNYYNDWYN--------EDLLELHELELDDLKTFYNKNKEIFDLYESRAELW 363
test6_dpinsect      SQQERKRFSNYYNKYYN--------EDLLELHELELDDLKSFYTCNKEIFDLYESRAELW 356
test12_sp           SEEQRKSFTPMYEDIIT--------EQVLTAHENYIKQLEAEVSANKSFLSLINRYASLI 388
test11_sc           SEESRSKFIIVFEELRNSATTLQEDELLLETCENELKRLEEKLTLYKPILKLISDFESLQ 540
                    *::.*. *     .  .        * :*   :  :  *.      * ::.  .   .

test2_hs            RLFLEFERKASDPNRFTNRG----GNLLKEEKQRAKLQKMLPKLEEELKARIELWEQEHS 418
test3_mm            KLFLEFERKASDPGRFTNRG----GNLLKEEKERAKLQKTLPKLEEELKARIEQWEQEHS 418
test4_dminsect      SRMEALEAKASEPNRFNNRG----GQLLKEEKERKAITSKLPKIEQQITELVQAYEAQEN 419
test6_dpinsect      SRMQALEAKANEPNRFNNRG----GQLLKEEKERKAISSKLPKIEQQITELVHAYEAQSH 412
test12_sp           EGKKELEASSNDASRLTQRGRRDPGLLLREEKIRKRLSRELPKVQSLLIPEITAWEERNG 448
test11_sc           EDQEFLERSSKDSSRLLSRN--SHKILLTEEKMRKRITRHFPRVINDLRIKLEEADGLFD 598
                         :* .:.:..*: .*.      ** *** *  :   :*:: . :   :   :
```

```
test2_hs            KAFMVNGQKFMEYVAEQWEMHRLEKERAKQERQLKN--------------------KKQ 457
test3_mm            TAFVVNGQKFMEYVTEQWELHRLEKERAKQERQLKN--------------------KKQ 457
test4_dminsect      TPFLVHGENILERMAADWERHRQAKQQSSARKKAPP-------------------SASKI 460
test6_dpinsect      GPFLVYGENILELMAGEWENYRQAKQ-SSARKKAPPTTRTG-------------SSSKLM 458
test12_sp           RTFLFYDEPLLKICQEATQPKSLYRSASAAANRPKTATTTDSVNRTPSQRGRVAVPSTPS 508
test11_sc           QPFLFKGKPLSEAIDIQQQEIEAKYPRCRVRMQRSKKGKCG--------------ANKEN 644
                     .*:. .: : :      :        .    :                        .

test2_hs            TETEMLYGSAPRTPSKRRGLAPNTPGKARKLNTTTMSNATA------NSSIRPIFGGTVY 511
test3_mm            TEAEMLYGSTPRTPSKRPG---QTPKKSGKMNTTTMSSATP------NSSIRPVFGGSVY 508
test4_dminsect      MPPPATGSTAPRTPRTLRNMSTLSSSTMSLRKTPSQQHLRPLNITKSTGNLHKRLNPSGL 520
test6_dpinsect      MPPPTAGSTAPRTPRTLKNMSSMSTSTMSLRKTPTIQLITP-NMTKSTGNLHKRLNPSAA 517
test12_sp           VRSASRAMTSPRTPLPRVKNTQNPSRSISAEPPSATSTANRRHPTANRIDINARLNSASR 568
test11_sc           KVIKNTFKATESSIRVPIGLNLNDANITYKTPSKKTIQGLTKNDLSQENSLARHMQGTTK 704
                           ::  :          .       .               .:   :  :
```

CLUSTAL W (1.82) multiple sequence alignment – 2nd Result

Hs (human/homo sapien), mm (mouse/mus musculus), dminsect (fruit fly1/drosophila melanogaster), dpinsect (fruit fly 2/drosophila probo), at (mustard weed/arabidopsis thaliana), sd (night shade/solanum demissum), sc (bakers yeast/Sacromyces cerevisiae), sp (brewers yeast/Sacromyces pombe)

```
test2_hs            ---VPSLEELNQFRQHVTTLRETKASRREEFVSIKRQIILCMEEL------------DH 190
test3_mm            ---VPTLEELKLFRQRVATLRETKESRREEFVNIKKQIILCMEEL------------EH 190
test4_dminsect      ---LPLPEEMDAFRNRLGQLRDQRVLRLKEMDQLRQAIKHDMKLL------------EC 193
test6_dpinsect      ---LPLPDEMDGFRDHLDNLRSQRAVRQTELDQLRKAIKHDMKML------------EL 186
test1_at            ---DLSLQRLEELRSQLGELQNEKSKRLEEVECLLKTLNSLCSVLGE----------DF 170
test7_sd            ---DLSLRKLEELHTELHTLQKEKSERLKQVLNHLGTLNSLCSVLGM----------DF 188
test11_sc           AFKTINEESVKHMNEVIKIYEEEYERRFKSVLTKKVSISSICEQLGTPLATLIGEDFEQD 360
test12_sp           ---DVSDAFTESLRGRINEAEKEIDARLEVINSFEEEILGLWSELGVEP---------AD 217
                              . :.  :   ..    *   .      :    . *

test2_hs            TPDTSFERDVVCEDEDAFCLSLENIATLQKLLRQLEMQKSQNEAVCEGLRTQIRELWDRL 250
test3_mm            SPDTSFERDVVCEDESAFCLSLENIATLQKLLKQLEMKKSQNEAECEGLRTQIRELWDRL 250
test4_dminsect      LPQTDTEERLLN--QVDHCLTPETFELLRNMQKNFADQVKELRERIDDMREKIHVLWDRL 251
test6_dpinsect      MPQTDAEDRLLN--DLSHNLTPETLERLRQMRNSYAEQIQELRSKIHDMREKIYVLWDRL 244
test1_at            KGMIRGIHSSLVDSN-TRDVSRSTLDKLDMMIVNLREAKLQRMQKVQDLAVSLLELWNLL 229
test7_sd            KHTINEVDPNLGESEEAKNICDDTIQNLAATIQRLQEVKLQRMQRLQDLTTSMLELWNLM 248
test11_sc           LRSYGEEENSTSEIPNFHPVDRERMSKIDITLEKLQAIHKERADKKRLLMEQCQKLWTRL 420
test12_sp           VPQYEQLLESHTNRPNDVYVTQELIDQLCKQKEVFSAEKEKRSDHLKSIQSEVSNLWNKL 277
                                     :  . :  :           :       :  .   **  :

test2_hs            QIPEEER-------------EAVATIMSGSKAKVRKALQLEVDRLEELKMQNMKKVIEAI 297
test3_mm            QIPEEER-------------EPVEAIMTGSKTKIRNALKLEVDRLEELKMQNIKQVIEKI 297
test4_dminsect      QETDEYA-------------MRRVREATTYTQRTYDVLREELQRCQALRRQNLKTFIEQL 298
test6_dpinsect      QETDESA-------------MRRVRECTENTQRTYDILHSELQRCQALRSQNLKTFIEQL 291
test1_at            DTPAEEQKIFHNVTCSIALTESEITEANILSVASIKRVEDEVIRLSKIKITKIKEVILRK 289
test7_sd            DTPIEEQQMFQNVTCKIAAKEHEITEPNMLSMEFITYVEEEVDRLEELKASKMKELVLKK 308
test11_sc           KISQEYI-------------KTFMRNNSSLSTESLGRISKEVMRLEAMKKKLIKKLISDS 467
test12_sp           QVSPNEQ-------------SQFGDSSNINQENISLWETELEKLHQLKKEHLPIFLEDC 323
                    . . :                       .  .         *: :   ::   :  .:

test2_hs            RVELVQYWDQCFYSQEQRQA------FAPFCAEDYTES--LLQLHDAEIVRLKNYYEVHK 349
test3_mm            RVELAQFWDQCFYSQEQRQA------FAPYYSEDYTEN--LLHLHDAEIVRLRNYYDAHK 349
test4_dminsect      RIEIKEMWDLTLKSQQERKR------FSNYYNDWYNED--LLELHELELDDLKTFYNKNK 350
test6_dpinsect      RVEISKWWDLTLKSQQERKR------FSNYYNKYYNED--LLELHELELDDLKSFYTCNK 343
test1_at            RLELEEISRKMHMATEVLKSENFSVEAIESG-VKDPEQ--LLEQIDSEIAKVKEEASSRK 346
test7_sd            KSELEEIYRKTHMVGDSDGAMNIAIEAIESGAVNDADA--VLEQIELRIAQVKEEAFSRK 366
test11_sc           WDKIQELWRTLQYSEESRSKFIIVFEELRNSATTLQEDELLLETCENELKRLEEKLTLYK 527
test12_sp           RQQILQLWDSLFYSEEQRKS------FTPMYEDIITEQ--VLTAHENYIKQLEAEVSANK 375
                      :: :         :                    :   :*   :  :  :.      *
```

```
test2_hs            ELFEGVQKWEETWRLFLEFERKASDPNRFTN---RGGNLLKEEKQRA-KLQKMLPKLEEE 405
test3_mm            ELFQGVQKWEESWKLFLEFERKASDPGRFTN---RGGNLLKEEKERA-KLQKTLPKLEEE 405
test4_dminsect      EIFDLYESRAELWSRMEALEAKASEPNRFNN---RGGQLLKEEKERK-AITSKLPKIEQQ 406
test6_dpinsect      EIFDLYESRAELWSRMQALEAKANEPNRFNN---RGGQLLKEEKERK-AISSKLPKIEQQ 399
test1_at            EILEKVEKWMSACEEESWLEEYNRDDNRYNAG--RGAHLTLKRAEKARLLVNKLPGMVEA 404
test7_sd            DILDRVEKWIAACEEECWLEEYNRDENRYNAG--RGTHLTLKRAEKARALVNKLPAMVEA 424
test11_sc           PILKLISDFESLQEDQEFLERSSKDSSRLLSR--NSHKILLTEEKMRKRITRHFPRVIND 585
test12_sp           SFLSLINRYASLIEGKKELEASSNDASRLTQRGRRDPGLLLREEKIRKRLSRELPKVQSL 435
                     ::.  .           :*    : .*      ..  :   . :    :   :* : .

test2_hs            LKARIELWEQEHSKAFMVNGQ---KFMEYVAEQWEMHRLEKERAKQERQLKN-------- 454
test3_mm            LKARIEQWEQEHSTAFVVNGQ---KFMEYVTEQWELHRLEKERAKQERQLKN-------- 454
test4_dminsect      ITELVQAYEAQENTPFLVHGE---NILERMAADWERHRQAKQQSSARKKAPP------SA 457
test6_dpinsect      ITELVHAYEAQSHGPFLVYGE---NILELMAGEWENYRQAKQ-SSARKKAPPTTRTGSSS 455
test1_at            LTAKVTAWENERGNEFLYDGVRVLSMLGQYKTVWEEKEHEKQRQRDMKKLHG-------- 456
test7_sd            LASKTKAWENERGTQFSYDGIPLLSMLEEYTILREEKELERKKQRDQKKLQG-------- 476
test11_sc           LRIKLEEADGLFDQPFLFKGKPLSEAIDIQQQEIEAKYPRCRVRMQRSKKGKCGANKENK 645
test12_sp           LIPEITAWEERNGRTFLFYDEPLLKICQEATQPKSLYRSASAAANRPKTATTTDSVN--R 493
                    :       :      *   .    .          .

test2_hs            KKQTETEMLYGSAPRTPS---KRRGLAPNTPGKARKLNTTTMSNATA------NSSIRPI 505
test3_mm            KKQTEAEMLYGSTPRTPS---KRPG---QTPKKSGKMNTTTMSSATP------NSSIRPV 502
test4_dminsect      SKIMPPPATGSTAPRTPR---TLRNMSTLSSSTMSLRKTPSQQHLRPLNITKSTGNLHKR 514
test6_dpinsect      KLMMPPPTAGSTAPRTPR---TLKNMSSMSTSTMSLRKTPTIQLITP-NMTKSTGNLHKR 511
test1_at            QLITEQEALYGSKPSPNK--SGKKPLRTPVN-AAMNRKLSLGGAMLHQS---LKH-EKAT 509
test7_sd            QLMAEQESLYGSKPSPMKNQSAKKGPKLSCGGAPSNRRLSLGGTMQQTCKTELPHSTKAT 536
test11_sc           VIKNTFKATESSIRVPIGLNLNDANITYKTPSKKTIQGLTKNDLSQENSLARHMQGTTKL 705
test12_sp           TPSQRGRVAVPSTPSVRSASRAMTSPRTPLPRVKNTQNPSRSISAEPPSATSTANRRHPT 553
                              :                                 .
```

The Psi-blast alignment result showing conservation in the MT binding region in the 2 plant species:

>gi|4506039|ref|NP_003972.1| protein regulator of cytokinesis 1 isoform 1 [Homo sapiens]
gi|13111935|gb|AAH03138.1| Protein regulator of cytokinesis 1, isoform 1 [Homo sapiens]
gi|2865521|gb|AAC02688.1| protein regulating cytokinesis 1; PRC1 [Homo sapiens]
Length = 620

Score = 238 bits (610), Expect = 5e-62
Identities = 190/191 (99%), Positives = 190/191 (99%)

```
Query: 1   KMQNMKKVIEAIRVELVQYWDQCFYSQEQRQAFAPFCAEDYTESLLQLHDAEIVRLKNYY 60
           KMQNMKKVIEAIRVELVQYWDQCFYSQEQRQAFAPFCAEDYTESLLQLHDAEIVRLKNYY
Sbjct: 286 KMQNMKKVIEAIRVELVQYWDQCFYSQEQRQAFAPFCAEDYTESLLQLHDAEIVRLKNYY 345

Query: 61  EVHKELFEGVQKWEETWRLFLEFERKASDPNRFTNRGGNLLKEEKQRAKLQKMLPKLEEE 120
           EVHKELFEGVQKWEETWRLFLEFERKASDPNRFTNRGGNLLKEEKQRAKLQKMLPKLEEE
Sbjct: 346 EVHKELFEGVQKWEETWRLFLEFERKASDPNRFTNRGGNLLKEEKQRAKLQKMLPKLEEE 405

Query: 121 LKARIELWEQEHSKAFMVNGQKVMEYVAEQWEMHRLEKERAKQERQLKNKKQTETEMLYG 180
           LKARIELWEQEHSKAFMVNGQK MEYVAEQWEMHRLEKERAKQERQLKNKKQTETEMLYG
Sbjct: 406 LKARIELWEQEHSKAFMVNGQKFMEYVAEQWEMHRLEKERAKQERQLKNKKQTETEMLYG 465

Query: 181 SAPRTPSKRRG 191
           SAPRTPSKRRG
Sbjct: 466 SAPRTPSKRRG 476
```

```
>gi|48209900|gb|AAT40494.1|     putative microtubule-associated protein [Solanum
demissum]
          Length = 730

 Score =  185 bits (471), Expect = 6e-46
 Identities = 47/200 (23%), Positives = 89/200 (44%), Gaps = 14/200 (7%)

Query: 1   KMQNMKKVIEAIRVELVQYWDQCFYSQEQRQAFAPFCAEDYT------ESLLQLHDAEIV 54
           K   MK+++   + EL + + +     +   A         +      +++L+  +  I
Sbjct: 297 KASKMKELVLKKKSELEEIYRKTHMVGDSDGAMNIAIEAIESGAVNDADAVLEQIELRIA 356

Query: 55  RLKNYYEVHKELFEGVQKWEETWRLFLEFERKASDPNRFT-NRGGN--LLKEEKQRAKLQ 111
           ++K      K++ + V+KW          E    D NR+   RG +  L + EK RA +
Sbjct: 357 QVKEEAFSRKDILDRVEKWIAACEEECWLEEYNRDENRYNAGRGTHLTLKRAEKARALV- 415

Query: 112 KMLPKLEEELKARIELWEQEHSKAFMVNGQKVMEYVAEQ---WEMHRLEKERAKQERQLK 168
             LP + E L ++ + WE E    F  +G  ++  + E     E   LE+++ + +++L+
Sbjct: 416 NKLPAMVEALASKTKAWENERGTQFSYDGIPLLSMLEEYTILREEKELERKKQRDQKKLQ 475

Query: 169 NKKQTETEMLYGSAPRTPSK 188
            +   E E LYGS P +P K
Sbjct: 476 GQLMAEQESLYGSKP-SPMK 494


>gi|9758202|dbj|BAB08676.1| unnamed protein product [Arabidopsis thaliana]
 gi|15242132|ref|NP_199973.1| microtubule associated protein (MAP65/ASE1) family
protein
           [Arabidopsis thaliana]
          Length = 707

 Score =  181 bits (460), Expect = 1e-44
 Identities = 52/202 (25%), Positives = 87/202 (43%), Gaps = 13/202 (6%)

Query: 1   KMQNMKKVIEAIRVELVQYWDQCFY-----SQEQRQAFAPFCAEDYTESLLQLHDAEIVR 55
           K   MK+++   R EL +   +         S   +   A          +L+  +  I +
Sbjct: 311 KASKMKELVLKKRSELEEICRKTHLLPVSDSAIDQTIVAIESGIVDATMVLEHLEQHISK 370

Query: 56  LKNYYEVHKELFEGVQKWEETWRLFLEFERKASDPNRFTNRGGN---LLKEEKQRAKLQK 112
           +K      KE+ E V+KW          E    D NR+    G    L + EK R  + K
Sbjct: 371 IKEEALSRKEILERVEKWLSACDEESWLEEYNRDDNRYNAGRGAHLTLKRAEKARNLVTK 430

Query: 113 MLPKLEEELKARIELWEQEHSKAFMVNGQKVMEYVAEQ---WEMHRLEKERAKQERQLKN 169
            LP + E L ++  +WEQE+   F+ +G +++  + E     +    E+  + +++L+
Sbjct: 431 -LPGMVEALASKTIVWEQENGIEFLYDGIRLLSMLEEYNILRQEREEEHRRQRDQKKLQG 489

Query: 170 KKQTETEMLYGSAPRTPSKRRG 191
           +   E E LYGS P +PSK  G
Sbjct: 490 QLIAEQEALYGSKP-SPSKPLG 510
```

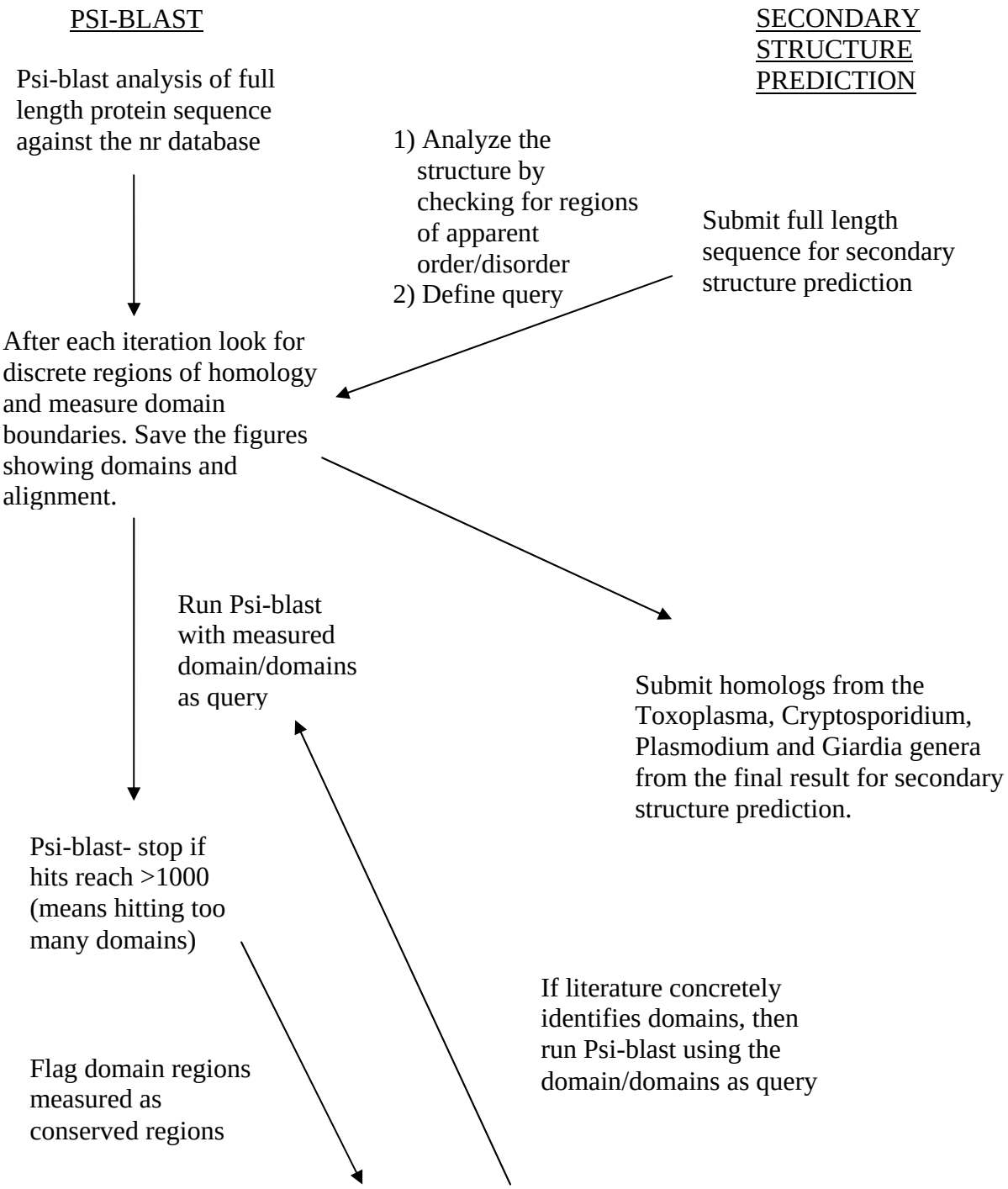
Figure 9: Comparison of Clustal W and Psi-blast result
The Psi-blast result shows that the conservation is there and the apparent loss of conservation in the Clustal W results was due to the limitations of its pair-wise alignment algorithm.

2.1.4. Discussion of the Limitations of Using Clustal W: The limitations of the Clustal W pair-wise alignment algorithm can be further elucidated by examining and comparing the principles

behind the Psi-Blast and Clustal W algorithms. Clustal W uses a progressive approach. The algorithm greedily adds sequences together, following the initial tree. There is no guarantee that the global optimal solution (defined by some overall measure of multiple alignment quality) will be found. More specifically, any mistakes (misaligned regions) made early in the alignment process cannot be corrected later as new information from other sequences is added. This problem is frequently thought of as resulting from an incorrect branching order in the initial tree. The initial trees are derived from a matrix of distances between separately aligned pairs of sequences (pair wise-alignment) and are much less reliable than trees from complete multiple alignments. The most demanding part of the multiple alignment strategy, in terms of computer processing and memory usage, is the alignment of two (groups of) sequences at each step in the final progressive alignment [9]. In contrast Psi-blast is a tool that produces a position-specific scoring matrix constructed from a multiple alignment of the top-scoring BLAST responses to a given query sequence. This scoring matrix produces a profile designed to identify the key positions of conserved amino acids within a motif. When a profile is used to search a database, it can often detect subtle relationships between proteins that are distant structural or functional homologues [9]. Thus Psi-blast provided an enormous advantage over Clustal W in the detection of distantly related sequences. For this reason the step to determine multiple sequence alignment using Clustal W in the first pass was eliminated in the current pass or the modified protocol. The Modified Protocol – Current Pass is shown in Flow-chart: 2 on the following page.

The Modified Protocol – Current Pass:

(Assumption – that within a microtubule binding protein family there are common MTB domains)



Scan literature for domain analysis and for microtubule binding or association on the query protein. Note down the method (assay/ immunoprecipitation/ extraction using tubulin etc), used to show MT binding or association and the results obtained

2.1.5. Secondary Structure Prediction, First Pass: Secondary structure prediction of the full length protein sequence was obtained using the PredictProtein tool. Within the results obtained using this predictor, the PROFsec and COILS program results were saved used in our analysis. The secondary structure was analyzed for transmembrane helix locations and topology and coiled coil regions; to basically flag conserved regions that appear to be ordered. In the case of Ensconsin (E-MAP-115-105) secondary structure analysis, two regions (60-165 aa) and (460-630 aa) appeared to be ordered.

The delineation of these regions was clear in the case of Ensconsin. This proved that secondary structure prediction of the protein sequence can throw light on and help flag conserved regions. But in the case of Hook (homolog 3), the secondary structure analysis did not help identify specific ordered regions. The sequence of the four homologs of the query protein (from the Psi-blast result, with homology ~70%), was submitted for secondary structure prediction. In the analysis of Ensconsin, PRC1 (Protein regulator of cytokinesis 1) and Hook (homolog 3), when the secondary structures of the homologs were compared with the secondary structure of the query protein, none of them showed a clear break in boundaries to show apparent regions of order or conserved regions across the board. Not much could be inferred from their result so this step (to do the secondary structures prediction analysis of the homolog sequences) was eliminated in the modified protocol. It was however decided that the step to submit homologs from the Toxoplasma, Cryptosporidium, Plasmodium and Giardia genera that are picked from the Psi-blast alignment results for secondary structure prediction must be retained. This is because these organisms are of pathogenic interest and it would be worthwhile if secondary structure prediction can help flag regions of conservation in proteins homologs from these organism when MAPs are analysed.

2.1.6. Literature Analysis, First Pass: The name of the protein was taken and a search was performed in PubMed, protein and nucleotide databases at http://ww.ncbi.nlm.nih.gov. The resulting

information was used to compile a list of articles to look at as well as to search for additional names for a given protein. This is important since the characterization may have been done on a single homolog of the family. Additionally it is still quite common for different groups to be working on the same protein but with different names. Sometimes entries at the protein and nucleotide level contain article references that do not get flagged by a PubMed search. One check was to rescan PubMed with each name for the protein. Basically the first question addressed was whether the protein has been shown to bind to MTs or only to be MT associated. This was our first area of mis-annotation to avoid. For example: For two proteins A and B consider the following possible scenarios of protein MT binding; 1) Both proteins A and B bind to MTs, 2) Protein B binds to MTs and protein A binds to protein B and 3) Protein B binds to MTs, undergoes a conformational change to shape B', protein A only binds to protein B'. So, it was very important to note down clearly how a literature reference characterizes the MT binding or MT association of a protein. MT association can be shown by any of these methods: immunofluoresence studies showing co-localization, electron microscopy showing co-localization, immunoprecipitation, co-localization in a spin down assay (not purified in lysate), protein pull down by microtubule binding, protein pull down by tubulin in lysate, tubulin pull down by protein in lysate, microtubule pull down by protein in lysate and co-localization during extraction of tubulin from cells. MT binding can be shown by co-localization in a spin down assay (in vitro purified system), protein pull down by microtubule binding, protein pull down by tubulin in vitro, tubulin pull down by protein in vitro, microtubule pull down by protein in vitro and the in vitro overlay assay. The second question addressed during literature research was whether the reference indicates a region/domain involved in MT binding. If domain analysis is shown, it was noted how this is indicated; by homology, a predicted domain with favorable pI or by truncational analysis. However, a loss of binding could be due to the loss of a domain or due to disordering of the protein. Paying attention to secondary structure analysis of the protein in literature and also going through a number of articles that analyze the protein for research helps give clearer answers. To

overcome the shortcomings that were experienced during the first-pass of the protein query through the analysis, changes were incorporated in the modified protocol. The steps of the current pass are briefly discussed here.

2.1.7. Psi-blast, Current Pass: Psi-blast analysis of the full length protein sequence was done against the nr (non-redundant database). Psi-blast iterations were stopped if the hits were greater than 1000. After each iteration, the distribution of blast hits with the query sequence was scanned for regions of homology. The figures showing homologous regions and alignment were studied. Four homologs of the query protein (from the Psi-blast results with homology ~ 70%), were picked for further analysis.

2.1.8. Secondary Structure Prediction, Current Pass: Secondary structure prediction of the full length protein sequence was obtained using the PredictProtein tool. Within the results obtained using this predictor, the PROFsec and COILS program results were saved and used in our analysis. The secondary structure was analyzed for transmembrane helix locations, topology, and coiled coil regions, to flag conserved regions that appear to be ordered.

2.1.9. Literature Analysis, Current Pass: Literature analysis was done as described in the initial protocol. For every protein, literature references flagged by PubMed were read and analyzed. Information such as, by which methods or how MT binding and MT association of the protein are shown and which region/domain is indicated in the MT binding or association was noted down. This information was later organized in a concise form to facilitate its entry into the MAP-DB database. If literature identified domains in the protein, Psi-blast was run with the measured domain/domains as query and the subsequent steps indicated in the modified analysis were followed

II.B. The Microtubule Binding and Associated Protein Database (MAP-DB)

II.B.1. Conceptual design of the Database

The first step in designing the database was to consider the nature of the project (a biological protein database in this case) and to think about what features and information the users (researchers and others in the scientific community) of such a database would want access to. The data model was the initial part of the conceptual design process. It focuses on what data should be stored in the database while a functional model later on dealt with how the data was processed. Planning and analysis preceded this. In this manner, the scope of the database was decided. First data objects and relationships were identified. An initial ER (Entity Relationship) diagram was drafted with entities and relationships and this ER diagram was further analyzed and refined. Key attributes and then non-key attributes were added to the initial ER diagram. The data model was validated through normalization and integrity rules were imposed upon the model. This was not strictly a linear process; refinement and validation of the diagram uncovered problems of missing information, which required to be factored in.

II.B.2. Tools Used to Create and Implement the Database

2.1. Oracle and SQL Plus

Oracle is a very powerful database management system based on the Relational Database Model, provided by Oracle Corporation along with fully integrated database application development and administration tools. It uses Structured Query Language (SQL) for database access and its own proprietary procedural language PL/SQL for application development along with Java programming support. Oracle is the back end of this application. An account was obtained with the Oracle server

at the research lab of Dr. Mahesh Merchant, Associate Professor, Laboratory Informatics Program School of Informatics at IUPUI. The Oracle 9i Client software was installed, configured and used to access the account on the Oracle server. Sql Plus was used to create the database, enforce integrity constraints and to insert, delete and manage data in the database. The data within the database was easily accessed and manipulated in this manner. Fundamentally, data definitions in SQL were used to create, and alter the descriptions of the tables (or relations) of the database. SQL systems were used to specify primary keys and referential integrity constraints. Basic SQL queries like the select, delete, insert or alter statements were used for inserting information into and retrieving information from the database. Oracle provides support for Java - today's most popular and productive programming language. It has a robust, integrated, and scalable Java VM within the server. This expands Oracle's support for Java into all tiers of applications, allowing Java programs to be deployed where they perform best, either in the client, server, or middle tier without recompiling or modifying the Java code.

2.2. Microsoft FrontPage 2003

FrontPage 2003 is a Microsoft product that provides the features, flexibility, and functionality to help build better Web sites. It includes the professional design, data, and publishing tools needed to create dynamic and innovative Web sites. New layout and graphics tools make it easier to design exactly what we have in mind. FrontPage design tools can be used to generate better code. Built-in scripting tools can be used for interactive results. It is useful to work with graphics from other applications, giving more control over how images are displayed and saved. It can be used to create

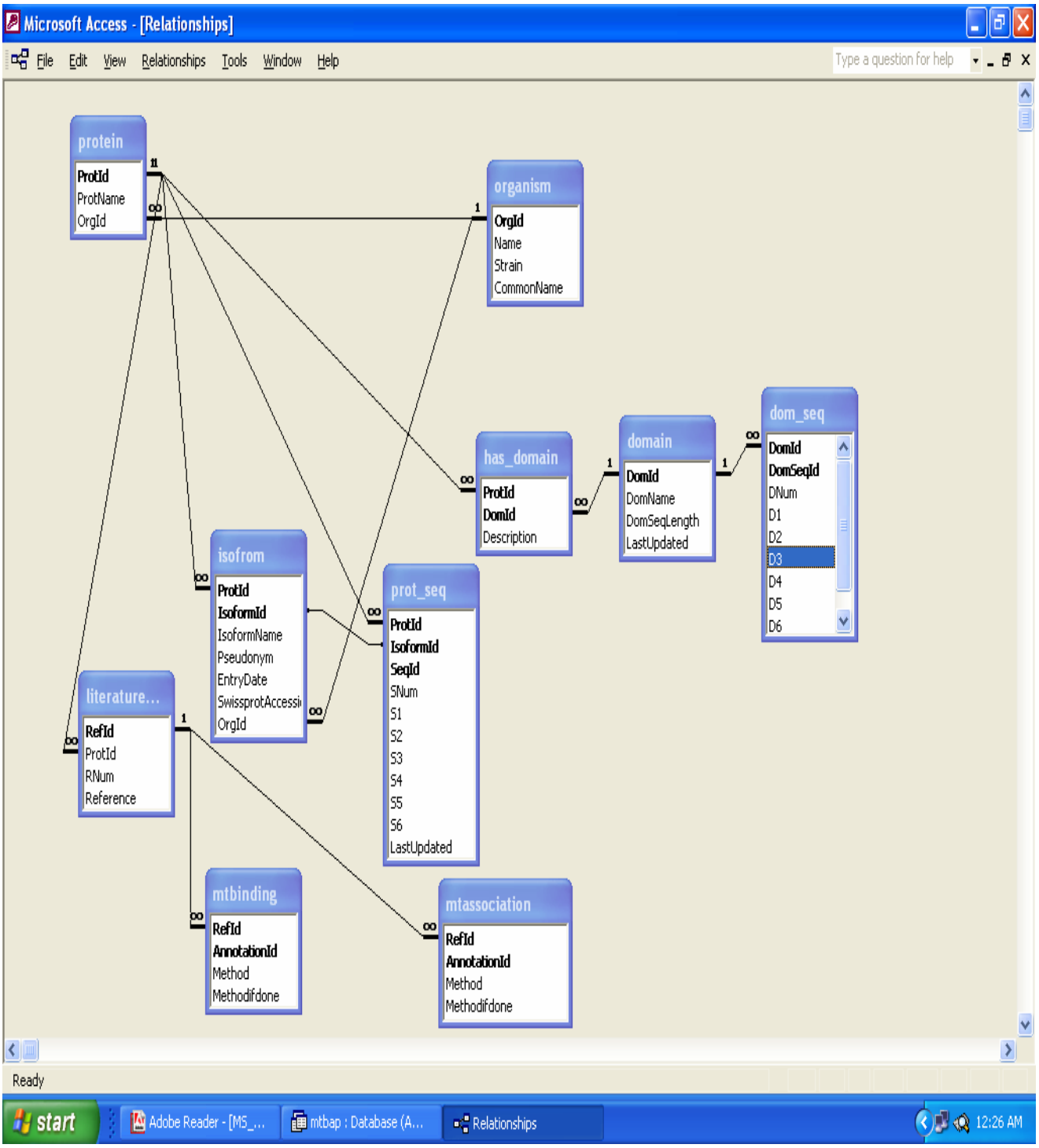


Figure 10: MS Access Screen-shot of the initial design of the database - the tables of the database (with their respective fields) and the integrity constraints imposed upon them.

and manipulate tables used for layout purposes, and it provides pixel-precise control of the layout [6]. Every file or web page created has the 'Design view' where the GUI is designed and a 'Code

view’ where HTML- Hypertext Markup Language code is inserted to manipulate the web page and devise its function. HTML is the coded format language used for creating hypertext documents on the World Wide Web and for controlling how the web pages appear. When designing the web pages, HTML form tags, presentation tags, table tags, text tags and script tags were manipulated to create tables, spaces, blank lines etc and also to adjust the height, width layout and content of the page and the tables in it. SQL code and Java Database Connectivity (JDBC) was used to pull data from the backend oracle database and JSP code was used to create dynamic web-content.

2.3. JSP and JDBC

Java Server Pages (JSP) and Java Database Connectivity (JDBC) technologies were used to integrate the static and dynamic database content in the Web site. With JSP technology, a developer can write static HTML pages and simply add Java code in those sections of the page that need to be dynamically generated. This flexibility enables rapid development of simple Web applications. It brings the "write once, run anywhere" paradigm to interactive Web pages. JSP lets the developer wield Java as a server-side scripting language. The biggest advantage of using JSP is that it helps effectively separate presentation from content. The JSP pages use short scriptlets to expose the code to its underlying JDBC concepts. Our source of dynamic content is the oracle relational database. Application of JSP technology to relational databases was done through JDBC. JDBC is the means by which Java programs work with relational databases. JDBC is the bridge between Java code and SQL databases. The primary JDBC objects represent connections to a database and the statements are performed using those connections. The two basic kinds of statements used with a relational database are queries and updates. As a prerequisite to each, it is first required that a connection to the database must be established, which is done with the java.sql.DriverManager class. The code snippet listed in the Figure: 11 below illustrates how to establish a connection with a test database, create a statement object to use that connection, issue an SQL query, process the results, and release the

JDBC resources. Along with JSP and JDBC code, HTML (**H**yper **T**ext **M**arkup **L**anguage) was also used.

```
Connection connection = DriverManager.getConnection(URL, user, password);
Statement statement = connection.createStatement();
ResultSet results = statement.executeQuery(sqlQuery);

while (results.next())
{
    ... process query results ...
    logSQLWarnings(results.getWarnings());
}

results.close();
statement.close();
connection.close();
```

Figure 11: JDBC logic

2.4. Installation and Configuration of the Apache Tomcat Web Server

Apache is a free, fully configurable Web server. It is one of the most popular servers available. A web server is basically software on a machine, that replies to data requests from a browser using the WWW protocol called HTTP, allowing the of access HTML and JSP files. Tomcat is the servlet container that is used in the official Reference Implementation for the Java Servlet and JavaServer Pages technologies [8]. However first step was to download and install Java. The servlet 2.4 (JSP 2.0) specifications require JDK 1.3 or later. Java Software Development Kit (JSDK) 1.4.2_05 was downloaded from the Sun java web-site. The JSDK has development tools (in the bin subdirectory.) that help develop, execute, debug, and document programs written in the Java programming language. It has the runtime environment (in the jre subdirectory.), an implementation of the Java 2 runtime environment for use by the JSDK. The runtime environment includes a Java virtual machine, class libraries, and other files that support the execution of programs written in the Java programming language. JSDK also has (in the lib subdirectory) - additional class libraries and support files required by the development tools. Jakarta-tomcat-4.0.1 was downloaded from the Apache Jakarta Project web site for the current release build of Tomcat. The installations needed to

be further configured. The JAVA_HOME environment variable was set to tell Tomcat where to find Java. Failing to properly set this variable prevents Tomcat from compiling JSP pages. The instructions and steps to do these configurations were obtained mainly from the Apache Jakarta Project web site and from an online tutorial on Core Servlets [7].

2.5. Web-site design and the Dynamics of the Web pages

Four screens were designed for MAP-DB web site. The first screen briefly introduces the web site and provides the user the option of searching for a particular microtubule binding or microtubule associated protein by entering its name or synonym. The second screen pulls from the database and lists, all the isoforms for the particular protein that the user searched for, on the first screen. Each isoform listed has a 'View details' hyper-link beside it. Upon clicking this hyper-link details of that isoform are displayed on the third screen. This third screen has the following information: under the heading 'Entry Information' the primary accession number, the corresponding Swiss Prot accession number and the date the entry was put into the database is listed. Under the heading 'Name and Origin of the Protein': the name of the protein, the synonym that it is commonly addressed by, which organism it belongs to and the domain/domains of interest in that protein are shown. Next, the protein sequence is displayed; the manner in which this sequence is displayed with numbering of its amino acids is explained in the next paragraph. Finally on the third screen the References used for annotation of the protein are listed. Against each reference, the methods used by that reference for determination of MT binding and MT association is listed in a tabular form (in two adjacent tables). If the "View Domain" hyper-link is clicked on the third screen, it takes the user to a fourth screen that displays details of the domain that was selected. On this screen the name, number and range of amino acids in that domain and the domain sequence are shown. In addition to this the page also lists the references and MT binding and MT associated methods used to annotate that domain. In this manner the database strives to eventually display meticulously annotated information that is

thoroughly corroborated by research and displayed clearly for all microtubule binding or microtubule associated proteins.

2.6. Elucidation of Some Key Features in the Database

A fundamental technique in JDBC was used to connect to the backend Oracle database, pull data from its relevant tables, and then display them on the web pages. The principal block of JSP code that was used several times to appropriately pull data from the database is shown below; in this particular example the protein ID field was pulled from the protein table and displayed on the screen:

```
// imports relevant packages in the java language
<%@ page import="java.util.*, java.lang.*, java.sql.*, java.lang.String,
oracle.jdbc.driver.OracleDriver" %>
   <%
// a connection is established to the oracle database using java methods
try {
      String first_str;
      boolean found=false;
      Class.forName("oracle.jdbc.driver.OracleDriver");
      DriverManager.registerDriver(new oracle.jdbc.driver.OracleDriver());
      PreparedStatement ps;
      Connection dbconn = DriverManager.getConnection("jdbc:oracle:thin:@dbserv.uits.indiana.edu:1521:oed1", "nshenoy", "nshenoy");
      System.out.println("got connection");
      String last_name=request.getParameter("name");
      System.out.println("Value is "+last_name);
      Statement s = dbconn.createStatement();

//The ResultSet object was used. A table of data representing a database result set is generated by executing a statement that queries the database A ResultSet object maintains a cursor pointing to its current row of data. Initially the cursor is positioned before the first row. The next method moves the cursor to the next row, and because it returns false when there are no more rows in the ResultSet object, it can be used in a while loop to iterate through the result set.

 ResultSet rs1l = s.executeQuery("SELECT protein.ProtId FROM protein WHERE protein.ProtName = '" + last_name + "'");

  while (rs1l.next()) {
      first_str=rs1l.getString(1);

      //out.println("<b>"+rs.getString(1)+"</b>" + " " + "<b>" + rs.getString(2)+"</b>" + "<br>");
      found=true;
```

```
    %>
//Embedded HTML code to display the parameter/ parameters pulled from the database
       <tr>
<td width="226" bgcolor="#00FFFF" bordercolor="#000000" bordercolorlight="#00FFFF" bordercolordark="#00FFFF">
<font size="5">View of Protein ID:</font></td>
<td bordercolorlight="#00FFFF" bordercolordark="#00FFFF" bordercolor="#000000" bgcolor="#00FFFF"> <%=first_str%></td>
</tr> <%
     System.out.println("First String is " + rs1l.getString(1));

    }
    System.out.println("The flag is "+found);
    if (found==false)
    {
    out.println("<b>" + " The Entered Name Does Not Exist In the Database " + "</b>" + "<br>");
    }// The resultset object and connection to the database is closed
    rs1l.close();
   s.close();
   dbconn.close();
  }// Errors in the execution of the code above are caught and displayed.
  catch (ClassNotFoundException e1) {
  System.out.println(e1.toString());
 }
  catch (SQLException e2) {
  System.out.println(e2.toString());
 }
  catch (Exception e3) {
   System.out.println(e3.toString());
 } %>
```

To get the protein sequence displayed in the format shown in Figure: 12, fields SNum, S1, S2, S3, S4, S5, and S6 were added to the prot_seq table. The prot_seq table has the protein sequence information. The idea was to split the entire amino acid sequence into sets of 10. So each field (S1 through S6) could hold one set of 10 aminoacids. The SNum field holds the number corresponding to the starting number of the amino-acids displayed in that row of the sequence. For the 1st row SNum has the value 1, 1-60 amino acids of the sequence are displayed in the first row. For the 2nd row SNum has the value 61, 61-120 amino acids of the sequence are displayed in the second row

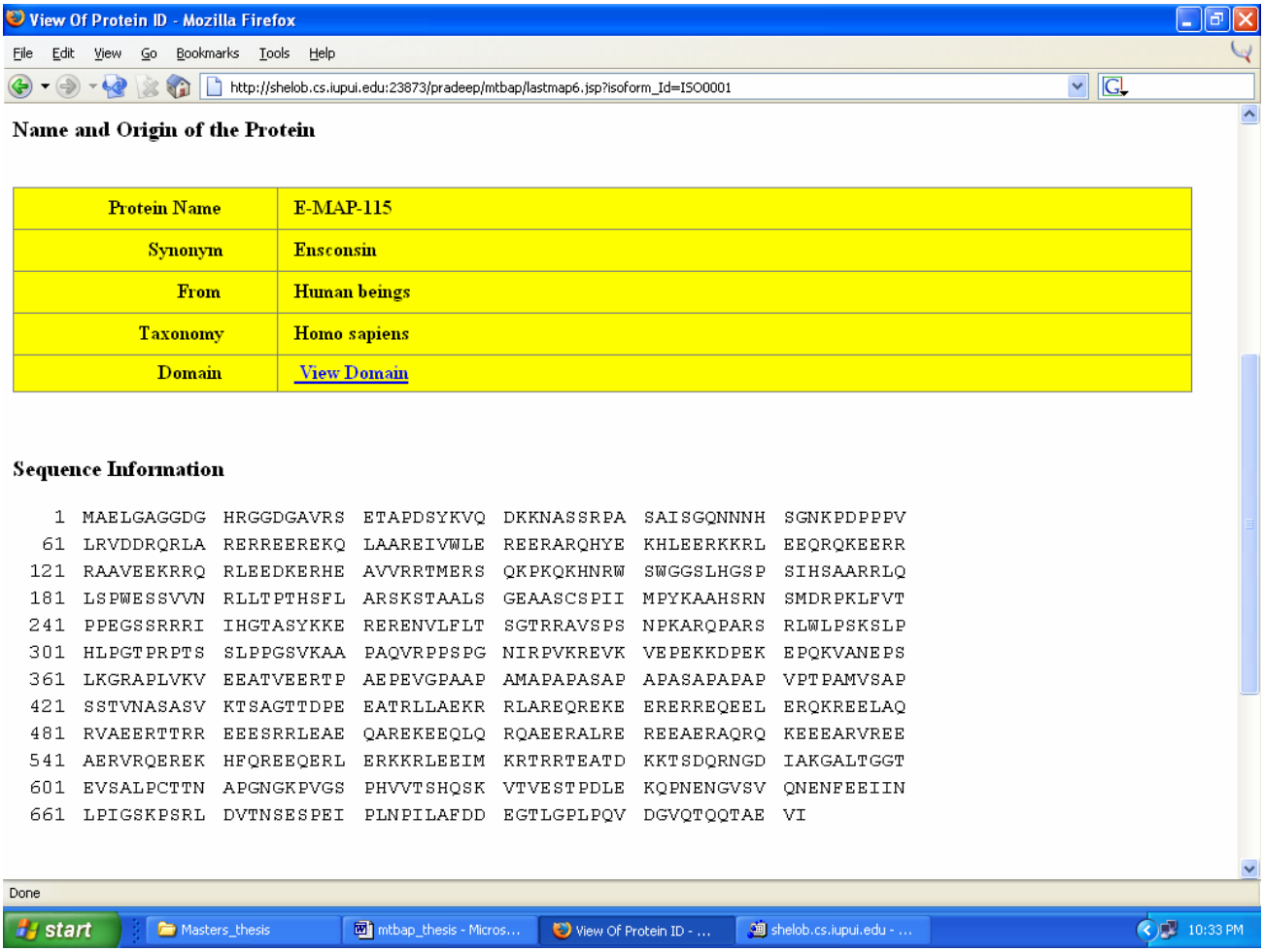


Figure 12: Format in which sequence information is displayed

and so on. An SQL select statement was used to pull these fields from the prot_seq table using a resultset object. The same technique was employed to display domain sequence information as well. Another key feature of the database is the description of references and annotations for a particular protein selected by the user. A reference table provides users data on the references that are studied for that particular protein. And against each reference, are two adjacent tables, one gives information on what methods were employed (if at all) in the reference to determine MT binding of the protein and the other gives information on what methods were employed to determine MT association of the protein. This information is well researched and is presented in a concise snapshot to the user. A similar design was implemented to present annotation of specific domains of the proteins in the database as well.

III. Results

III. A. Analysis and Annotation

The result of the analysis of two proteins, epithelial microtubule-associated protein (E-MAP-115-105) from Humans (also known as Ensconsin) and HOOK3 (Human Hook homolog 3) is outlined here.

III.A.1. Analysis of Ensconsin (E-MAP-115-105):

1.1. Psi-blast: This protein was analyzed through the first pass (the initial analysis protocol was followed). Psi-blast analysis of the full-length protein sequence was done. The Psi-blast was run against the nr (non redundant) database and the iterations were stopped once the hits reached 2000. The diagram (Figure: 13) depicting the distribution of the blast hits on the query sequence was analyzed. Discrete regions of homology were identified. Four homologs that showed nearly 70% homology with the query sequence were picked for further analysis based upon the discrete regions of homology. The profiles of the four homologs are shown in (Figure: 13):

Psi-blast - Sequences producing significant alignments:

1) >gi|37725922|gb|AAO38039.1| reticulocyte binding-like protein 2b [Plasmodium reichenowi]

```
Score = 86.0 bits (212), Expect = 4e-15
 Identities = 15/89 (16%), Positives = 52/89 (58%), Gaps = 1/89 (1%)

Query: 61   LRVDDRQRLARERREEREKQLAAREIVWLEREERARQHYEKHLEERKKRLEEQRQKEERR 120
            L+  ++++ A+ ++EE  K+    +   L++EE  ++  E+  + + ++ EE +++E+ +
Sbjct: 2767 LKRQEQEKQAQLQKEEELKRQEQEKQAQLQKEEELKRQ-EQEKQAQLQKEEELKRQEQEK 2825

Query: 121  RAAVEEKRRQRLEEDKERHEAVVRRTMER 149
            +A ++++   + +E +++ +      ++R
Sbjct: 2826 QAQLQKEEELKRQEQEKQAQLQKEEELKR 2854
```

2) gi|49037438|gb|AAT49026.1| IgA1 protease [Neisseria meningitidis]

```
Score = 54.4 bits (130), Expect = 1e-05
Identities = 55/399 (13%), Positives = 116/399 (29%), Gaps = 10/399 (2%)

Query: 39  PASAISGQNNNHSGNKPDPPPVLR--VDDRQRLARERREEREKQLAAREIVWLEREERAR 96
           P S  + Q       +     V R   +  ++ A   +++ E +  ARE+   ++ E+ R
Sbjct: 1   PPSPQANQAEEAKRQQAKAEQVKRQQAEAERKSAELAKQKAEAEREARELATRQKAEQER 60

Query: 97  QHYE---KHLEERKKRLEEQRQKEERRRAAVEEKRRQRLEEDKERHEAVV--RRTMERSQ 151
              E   +H +ER+      +QK E  R A     R++ E ++ + +A    RR  +  +
Sbjct: 61  SSAELARRHEKEREAAELSAKQKVEAEREAQALAVRRKAEAEEAKRQAAELARRHEKERE 120

Query: 152 KPKQKHNRWSWGGSLHGSPSIHSAARRLQLSPWESSVVNRLLTPTHSFLARSKSTAALSG 211
             +    +          +       R+ + +P +    +           RS     +
Sbjct: 121 AAELSAKQRVGEEERRQTAQSQPQRRKRRAAPQDYMAAS--QDRPKRRGHRSVQQNNVEI 178

Query: 212 EAASCSPIIMPYKAAHSRNSMDRPKLFVTPPEGSSRRRIIHGTASYKKERERENVLFLTS 271
             A    +    +   + + +      E  +        A   K +  E      +
Sbjct: 179 AQAQAELVRRQQEERKAAELLAKQRAEA-EREAQALAARRKAEAEEAKRQAAELAHRQEA 237

Query: 272 GTRRAVSPSNPKARQPARSRLWLPSKSLPHLPGTPRPTSSLPPGSVKAAPAQVRPPSPGN 331
             + A   +N KA   A++      K+L       R  + L          +    +
Sbjct: 238 ERKAAELSANQKAAAEAQALAARQQKALARQQEEARKAAELAVKQKAETERKTAELAKQR 297

Query: 332 IRPVKREVKVEPEKKDPEKEPQKVANEPSLKGRAPLVKVEEATVEERTPAEPEVGPAAPA 391
                + + E  +       Q+   + +        + +    E       E      A
Sbjct: 298 AAAEAAKRQQEARQTAELARRQEAERQAAELSAKQKAETDREAAESAKRKAEEEEHRQAA 357

Query: 392 MAPAPASAPAPASAPAPAPVPTPAMVSAPSSTVNASASV 430
           +         A     A             ST+ A  S
Sbjct: 358 QSQPQRRKRRAAPQDYMAASQNRPKRRGRRSTLPAPPSP 396
```

3) gi|22086284|gb|AAM90625.1| chimeric erythrocyte-binding protein MAEBL [Plasmodium falciparum]

```
Score = 52.1 bits (124), Expect = 6e-05
Identities = 35/132 (26%), Positives = 50/132 (37%), Gaps = 22/132 (16%)

Query: 41   SAISGQNNNHSGNKP---DPPPVLRVDDRQRLARERR-EEREKQLAAREIVWLEREERAR 96
              +G+    S  K          R+++ +R    RR EE  +   AR +    R E AR
Sbjct: 1089 KTETGRIEEESKKKEAMKRAEDARRIEEARRAEDARRIEEARRAEDARRVEIARRVEDAR 1148

 Query: 97   QHYEKHLEERKKRLEEQRQKEERRRAAVE-------------EKRRQRLE----EDKER 138
             +       E  KR+E  R+  E RRA +                E  R+  E    ED++R
Sbjct: 1149 RIEISRRAEDAKRIEAARRAIEVRRAELRKAEDARRIEAARRYENERRIEEARRYEDEKR 1208

Query: 139  HEAVVRRTMERS 150
              EAV R    R
Sbjct: 1209 IEAVKRAEEVRK 1220
```

4) gi|9910111|gb|AAF68025.2| myosin X [Homo sapiens]

```
Score = 50.9 bits (121), Expect = 1e-04
 Identities = 20/92 (21%), Positives = 41/92 (44%), Gaps = 10/92 (10%)

Query: 66  RQRLARERREER-----EKQLAAREIVWLEREERARQHYEKHLEERKKRLEEQRQKEERR 120
           Q+    R+++      + Q  A     LE+++  +Q  E  L   K+  + QR KE++
Sbjct: 850 AQQEEETRKQQELEALQKSQKEAELTRELEKQKENKQVEE-ILRLEKEIEDLQRMKEQQE 908

Query: 121 ----RAAVEEKRRQRLEEDKERHEAVVRRTME 148
               A++++ + +R +E +   E   R   E
Sbjct: 909 LSLTEASLQKLQERRDQELRRLEEEACRAAQE 940
```

Figure 13 : Ensconsin, Psi-blast result

1.2. Secondary Structure Prediction: Secondary structure prediction of full length Ensconsin E-MAP-115-105 was done using PredictProtein available at http://www.predictprotein.org/. The results of this prediction are shown in Figure: 14.

Ensconsin: Secondary Structure prediction:

```
            ...,....1....,....2....,....3....,....4....,....5....,....6
AA         |MAELGAGGDGHRGGDGAVRSETAPDSYKVQDKKNASSRPASAISGQNNNHSGNKPDPPPV|
PROF_sec   | EEE                                      HHHH               |

           ...,....7....,....8....,....9....,....10.1.,....11.1.,....12.1
AA         |LRVDDRQRLARERREEREKQLAAREIVWLEREERARQHYEKHLEERKKRLEEQRQKEERR|
PROF_sec   |    HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH|

           ...,....13.1.,....14.1.,....15.1.,....16.1.,....17.1.,....18.1
AA         |RAAVEEKRRQRLEEDKERHEAVVRRTMERSQKPKQKHNRWSWGGSLHGSPSIHSADPDRR|
PROF_sec   |HHHHHHHHHHHHHHHHHHHHHHHHHHHHH                               |

           ...,....19.1.,....20.1.,....21.1.,....22.1.,....23.1.,....24.1
AA         |SVSTMNLSKYVDPVISKRLSSSSATLLNSPDRARRLQLSPWESSVVNRLLTPTHSFLARS|
PROF_sec   |                                                            |

           ...,....25.1.,....26.1.,....27.1.,....28.1.,....29.1.,....30.1
AA         |KSTAALSGEAASCSPIIMPYKAAHSRNSMDRPKLFVTPPEGSSRRRIIHGTASYKKERER|
PROF_sec   |                                                            |

           ...,....31.1.,....32.1.,....33.1.,....34.1.,....35.1.,....36.1
AA         |ENVLFLTSGTRRAVSPSNPKARQPARSRLWLPSKSLPHLPGTPRPTSSLPPGSVKAAPAQ|
PROF_sec   |                                                            |


           ..,....37.1.,....38.1.,....39.1.,....40.1.,....41.1.,....42.1
AA         |VRPPSPGNIRPVKREVKVEPEKKDPEKEPQKVANEPSLKGRAPLVKVEEATVEERTPAEP|
PROF_sec   |                                          EEEEE  EE         |
```

```
          ...,....43.1.,....44.1.,....45.1.,....46.1.,....47.1.,....48.1
AA        |EVGPAAPAMAPAPASAPAPASAPAPAPVPTPAMVSAPSSTVNASASVKTSAGTTDPEEAT|
PROF_sec |                                  EE      EEEEEEEE       HHHHH|

          ...,....49.1.,....50.1.,....51.1.,....52.1.,....53.1.,....54.1
AA        |RLLAEKRRLAREQREKEERERREQEELERQKREELAQRVAEERTTRREEESRRLEAEQAR|
PROF_sec |HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH    HHHHHHHHHHHH|

      ....,....55.1.,....56.1.,....57.1.,....58.1.,....59.1.,....60.1
AA        |EKEEQLQRQAEERALREREEAERAQRQKEEEARVREEAERVRQEREKHFQREEQERLERK|
PROF_sec |HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH|

          ...,....61.1.,....62.1.,....63.1.,....64.1.,....65.1.,....66.1
AA        |KRLEEIMKRTRRTEATDKKTSDQRNGDIAKGALTGGTEVSALPCTTNAPGNGKPVGSPHV|
PROF_sec |HHHHHHHHHH                                                  E|

          ...,....67.1.,....68.1.,....69.1.,....70.1.,....71.1.,....72.1
AA        |VTSHQSKVTVESTPDLEKQPNENGVSVQNENFEEIINLPIGSKPSRLDVTNSESPEIPLN|
PROF_sec |E                       EEE      EEEE                        |

          ...,....73.1.,....74.1.,....75.1
AA        |PILAFDDEGTLGPLPQVDGVQTQQTAEVI|
PROF_sec | EE                     EEE  |
```

Figure 14: Ensconsin, Secondary Structure prediction

1.3 Analysis of Secondary Structure Prediction:

From the secondary structure prediction of the full-length protein, it was inferred that only the (60-165 aa) and (460-630 aa) regions of E-MAP-115-105 appeared to be ordered. The conserved region (60-165 aa) from the secondary structure prediction result was flagged as a conserved region/ domain because it is a basic predicted alpha helical region that is highly charged with a positive pI value and thus more likely to bind the negatively charged surface of MTs.

1.4. Clustal W

From the Psi-blast result the regions of the homologs that aligned with the (60-165 aa) region of the query protein were further analyzed for multiple sequence alignment using Clustal W. Based on the regions of homology of each protein with the query sequence from the psi-blast results, the following amino acid regions of the query sequence and these homologs were demarcated for multiple sequence alignment using Clustal W:

1) Seq1-query E-MAP-115-105 (1-160 aa)

2) Seq2- reticulocyte binding-like protein 2b [Plasmodium reichenowi] (2767-2854 aa)

3) Seq3- IgA1 protease [Neisseria meningitidis] (1-396 aa)

4) Seq4- chimeric erythrocyte-binding protein MAEBL [Plasmodium falciparum] (1089-1220 aa)

5) Seq5- myosin X (850-940 aa)

The Clustal W Result:

```
Seq2            ----------------------------LKRQEQEKQAQLQKEEELKRQEQEKQAQLQKE- 32
Seq3            ERKAAELSANQKAAAEAQALAARQQKALARQQEEARKAAELAVKQKAETERKTAELAKQR 297
Seq5            ----------------------------AQQEEETRKQQELEALQKSQ---KEAELTREL 29
Seq4            ----------------------KTETGRIEEESKKKEAMKRAEDARRIEEARRAEDARRI 38
Seq1            ---MAELGAGGDGHRGGDGAVRSETAPDSYKVQDKKNASSRPASAISGQNNNHSGNKPDP 57
                                        : .. :   .           . :

Seq2            ---EELKRQ-EQEK--QAQLQKEEELKRQEQEKQAQLQKEEE-------LKRQEQEKQAQ 79
Seq3            AAAEAAKRQQEARQ--TAELARRQEAERQAAELSAKQKAETDREAAESAKRKAEEEEHRQ 355
Seq5            ---EKQKENKQVEE--ILRLEKEIEDLQRMKEQQELSLTEAS-------LQKLQERRDQE 77
Seq4            EEARRAEDARRVEIARRVEDARRIEISRRAEDAKRIEAARRAIEVRRAELRKAEDARRIE 98
Seq1            PPVLRVDDRQRLARERREEREKQLAAREIVWLEREERARQHYEKHLEERKKRLEEQRQKE 117
                      .   .       .  :.    .           .          :: :: .  :

Seq2            LQKEEELKR---------------------------------- 88
Seq3            AAQSQPQRRKRRAAPQDYMAASQNRPKRRGRRSTLPAPPSP-- 396
Seq5            LRRLEEEACRAAQE----------------------------- 91
Seq4            AARRYENERRIEEARRYEDEKRIEAVKRAEEVRK--------- 132
Seq1            ERRRAAVEEKRRQRLEEDKERHEAVVRRTMERSQKPKQKHNRW 160
```

Figure 15: Ensconsin, Clustal W Result

1.5. Analysis of Clustal W Result:

This is demonstrated by comparing the highlighted regions (Figure: 15) of the multiple sequence alignment of Seq1 (E-MAP_115-105, the query) with Seq4 (chimeric erythrocyte-binding protein MAEBL) and Seq5 (myosin X) in the clustal W result, with the highlighted region (shown in bold in Figure: 15) of alignment of the query with chimeric erythrocyte-binding protein MAEBL and myosin X in the Psi-blast result. On close examination, it was evident that the residues of these two

homologs aligned with the query in the Clustal W result differ significantly from the residues of these homologs that are aligned with the same region of the query in the Psi-blast result. This was due to the limitations of the Clustal W pair-wise alignment algorithm. Any mistakes (misaligned regions) made early in the alignment process in Clustal W cannot be corrected later as new information from other sequences is added. The step to determine multiple sequence alignment using Clustal W in the first pass was eliminated in the current pass or the modified protocol.

1.6. Literature Analysis:

The name of the protein was taken and a search was given using keywords “E-MAP-115 MT binding” and “E-MAP-115 MT associated” in PubMed, protein and nucleotide databases at http://ww.ncbi.nlm.nih.gov. The resulting information was used to compile a list of papers. Each paper was read and analyzed. Information such as, by which methods or how MT binding and MT association of the protein are shown and which region/domain is indicated in the MT binding or association was noted down. Details of two key references analyzed for Ensconsin are listed in Table: 1 on the following page.

Paper Reference	MT Binding/MT Association	Region/Domain Indicated in MT Binding or MT Association
1. Bulinski and Bossler Journal of Cell Science 1994	•Co-localization during extraction of tubulin from HeLa cells. •Immunoblotting assays of ensconsin using Coomassic Blue stain. Electrophoretic transfer and analysis of western blots with antibodies against ensconsin, tubulin and MAP4. •Immunofluorescence	•It is likely that Ensconsin possesses one or more MT binding sites structurally distinct from the MT binding sites of other MAPs like Tau, MAP4 and MAP2
2. D Masson and T E Kreis, The Journal of Cell Biology, 1993	•Sequence Analysis and Comparison by searching EMBL and Swiss Prot data libraries with BLAST and FASTA programs •Secondary structure and charge display of E-MAP-115 were predicted using the peptide structure program •Co-localization in an in-vitro assay by transfection into cells using mutated cDNA. Truncation analysis of E-MAP-115	•Tau, MAP1B, MAP2, MAP-U, 205 KD Drosophila MAP, kinesin heavy chain from Squid, CLIP-170 and dynamin share no homology with E-MAP-115 •E-MAP-115 can be divided into 6 main domains- a nearly neutral 59 aa amino terminal domain, a basic predicted alpha helical aa 60-149 region, a poorly structured basic aa 150-416 region, an alanine/praline rich region (PAPA, 417-457 aa), a highly charged second predicted alpha helical region (aa 477-619) and an acidic COOH terminus (aa 620-749) •Microtubule binding site of E-MAP-115 is in the amino terminal basic domain with the highly charged predicted alpha helical structure (aa 60-149)

Table:1

The literature analysis of Ensconsin (E-MAP-115-105) identified the basic N-terminal alpha helical (60-149 aa) region as the microtubule binding domain. Totally eight references were compiled and analyzed for Ensconsin.

III.A.2. Analysis of Hook (homolog 3):

2.1. First Pass: This protein was analyzed initially through the first pass (the initial protocol). Psi-blast analysis of the full-length protein sequence was done. The Psi-blast was run against the nr (non redundant) database and the iterations were stopped once the hits reached 2000. The diagram (Fig XYZ) depicting the distribution of the blast hits on the query sequence was analyzed. Four homologs that showed nearly 70% homology with the query sequence were picked for further analysis based upon the discrete regions of homology. In the secondary structure prediction of full length protein, no clear boundaries for ordered regions were identified. When the secondary structures of the homologs were compared with the secondary structure of the query protein, none of the homologs structures showed a clear break in boundaries to show apparent regions of order or conserved regions across the board. Not much could be inferred from this result so this step (to do the secondary structures prediction analysis of the homolog sequences) in the initial protocol was eliminated. Multiple sequence alignment using Clustal W did not help flag conserved regions or domains in the query sequence and the result was inconclusive. So, this step in the initial protocol was eliminated.

2.2. Current Pass: In the current pass of analysis of Hook (homolog 3), literature analysis indicated that the N-terminal domain (1-164 aa) of the protein, bound directly to microtubules. Using 1-165 aa of the protein as the query sequence, Psi-blast iterations were run until the hits reached 1000 (to avoid hitting too many domains). Four homologs that showed nearly 70% homology with the query

sequence were picked for further analysis based upon discrete regions of homology. The profiles fro the Psi-Blast result and the homologs are shown in Figure: 16.

Psi-blast - Sequences producing significant alignments:

```
1) >gi|56966906|pdb|1WIX|A    Chain A, The Solution Structure Of Rsgi Ruh-026,
Conserved Domain Of Hook1 Protein From Mouse
          Length = 164

 Score =  249 bits (637), Expect = 2e-65
 Identities = 103/154 (66%), Positives = 125/154 (81%)

Query: 12  LCESLLTWIQTFNVDAPCQTVEDLTNGVVMAQVLQKIDPAYFDENWLNRIKTEVGDNWRL 71
           LC+SL+ W+QTF   +PCQ V+ LTNGV MAQVL +ID A+F E+WL+RIK +VGDNWR+
Sbjct: 10  LCDSLIIWLQTFKTASPCQDVKQLTNGVTMAQVLHQIDVAWFSESWLSRIKDDVGDNWRI 69

Query: 72  KISNLKKILKGILDYNHEILGQQINDFTLPDVNLIGEHSDAAELGRMLQLILGCAVNCEQ 131
           K SNLKK+L GI  Y HE LGQQI++  +PD+N I E +D  ELGR+LQLILGCAVNCE+
Sbjct: 70  KASNLKKVLHGITSYYHEFLGQQISEELIPDLNQITECADPVELGRLLQLILGCAVNCEK 129

Query: 132 KQEYIQAIMMMEESVQHVVMTAIQELMSKESPVS 165
           KQE+I+ IM +EESVQHVVMTAIQELMSK  P S
Sbjct: 130 KQEHIKNIMTLEESVQHVVMTAIQELMSKSGPSS 163



2) gi|55235145|gb|EAA14793.2|   ENSANGP00000016709 [Anopheles gambiae str. PEST]

          Length = 676

 Score =  198 bits (505), Expect = 5e-50
 Identities = 71/151 (47%), Positives = 106/151 (70%)

Query: 8   ERAELCESLLTWIQTFNVDAPCQTVEDLTNGVVMAQVLQKIDPAYFDENWLNRIKTEVGD 67
           ++ E+ ESL+ W+   N+ AP  TV++L++G  +AQ L +I P  F ++WL++IK++VG
Sbjct: 4   DKMEIYESLIRWLSELNLSAPHGTVQELSDGAALAQALNQIAPEVFTDSWLSKIKSDVGA 63

Query: 68  NWRLKISNLKKILKGILDYNHEILGQQINDFTLPDVNLIGEHSDAAELGRMLQLILGCAV 127
           NWRLK+SNL+KI++GI  Y  + L   +++   PD   I E  D  ELGR+LQLILGCAV
Sbjct: 64  NWRLKVSNLRKIIEGIYVYYQDELSLNLSEELRPDALKIAEKGDPHELGRLLQLILGCAV 123

Query: 128 NCEQKQEYIQAIMMMEESVQHVVMTAIQELM 158
           NC +KQ+YI  IM +EES+Q  +M A+Q++
Sbjct: 124 NCLEKQKYITQIMELEESLQRNIMAALQDIE 154
```

```
3) gi|54644710|gb|EAL33450.1|   GA10469-PA [Drosophila pseudoobscura]

          Length = 677

 Score =  206 bits (525), Expect = 2e-52
 Identities = 74/158 (46%), Positives = 102/158 (64%), Gaps = 1/158 (0%)

Query: 6   SLERAELCESLLTWIQTFNVDAPCQTVEDLTNGVVMAQVLQKIDPAYFDENWLNRIK-TE 64
           S  + E+  SLL W +T N++AP    E L +GV +AQ L +  P  F ++WL +IK +
Sbjct: 2   SAAKNEMYYSLLEWFKTLNLNAPHADAESLADGVAVAQALNQFAPESFTDSWLAKIKASA 61

Query: 65  VGDNWRLKISNLKKILKGILDYNHEILGQQINDFTLPDVNLIGEHSDAAELGRMLQLILG 124
           VG NWRL++SNLKK+ + + DY  E+L   ++DF  PDV  I E  D  EL R+LQL+LG
Sbjct: 62  VGINWRLRMSNLKKVTQSLYDYYSEVLNYTLSDFVKPDVQRIAEKCDLVELERLLQLVLG 121

Query: 125 CAVNCEQKQEYIQAIMMMEESVQHVVMTAIQELMSKES 162
           CAVNC +KQ YI  IM +EE +Q  +M A+QEL S  +
Sbjct: 122 CAVNCAKKQSYITEIMCLEEELQANIMRALQELESSRN 159


4) gi|38037645|gb|AAR08446.1|   DVL-binding protein DAPLE [Mus musculus]

          Length = 2009

 Score =  201 bits (513), Expect = 5e-51
 Identities = 52/171 (30%), Positives = 89/171 (52%), Gaps = 14/171 (8%)

Query: 3   SVESLERAELCESLLTWIQTFNVDAPCQT-----VEDLTNGVVMAQVLQKIDPAYFDENW 57
           +V  L    L   L+TW++TF               DL +G+ + Q++ +IDP   ++
Sbjct: 4   TVSQLVELFLQSPLVTWVKTFGSFGSGHQDNLTLYMDLVDGIFLNQIMLQIDPRPSNQ-- 61

Query: 58  LNRIKTEVGDNWRLKISNLKKILKGILDYNHEILGQQINDFTLPDVNLIGEHS----DAA 113
             RI   V ++  L+I NL  +++ I  Y  E+L QQ+    LP+V +IG+
Sbjct: 62  --RINKHVNNDVNLRIQNLSILVRNIKTYYQEVL-QQLIVMNLPNVLMIGKDPLSGKSME 118

Query: 114 ELGRMLQLILGCAVNCEQKQEYIQAIMMMEESVQHVVMTAIQELMSKESPV 164
           E+ ++L L+LGCAV CE+K+E+I+ I  ++   Q  ++  IQE+   +  V
Sbjct: 119 EIKKVLLLVLGCAVQCERKEEFIERIKQLDIETQAGIVAHIQEVTHNQENV 169
```

Figure 16: Hook (homolog 3), Psi-blast result

The Psi-Blast sequence alignment of these homologs with the (1-165 aa) region of the query sequence indicated a significant number of identities and positives residues. This indicted that the (1-165 aa) region of the Hook protein in humans shows conservation with the aligned homologs in mouse and mosquito. The secondary structure prediction of the (1-165 aa) region of Hook (homolog 3) is shown in Figure: 17.

Hook (homolog 3) - 1-165 aa region: Secondary Structure prediction:

```
          AA        |MFSVESLERAELCESLLTWIQTFNVDAPCQTVEDLTNGVVMAQVLQKIDPAYFDENWLNR|
          OBS_sec   |                                                            |
          PROF_sec  |     HHHHHHHHHHHHHHHHH          HHHHH HHHHHHHHHH         HHHHH|
         |
....,....7....,....8....,....9....,....10.1.,....11.1.,....12.1
          AA        |IKTEVGDNWRLKISNLKKILKGILDYNHEILGQQINDFTLPDVNLIGEHSDAAELGRMLQ|
          OBS_sec   |                                                            |
          PROF_sec  |HH      HHHHHHHHHHHHHHHHHHHHHHHH            EEE     HHHHHHHHH|

                     ....,....13.1.,....14.1.,....15.1.,....16.1.,....17.1
          AA        |LILGCAVNCEQKQEYIQAIMMMEESVQHVVMTAIQELMSKESPVS|
          OBS_sec   |                                             |
          PROF_sec  |HHHH       HHHHHHHHHH  HHHHHHHHHHHHHHHH      |
```

Figure 17: Hook (homolog 3), Clustal W result

The secondary structure analysis of this region of the Hook (homolog 3) protein did not indicate specific ordered region/regions. Literature analysis however concretely identified the (1-165 aa) region of Hook (homolog 3) as microtubule binding. Details of two key references analyzed for Hook (homolog 3) is shown in Table: 2 on the following page. Totally six references were compiled and analyzed for Hook (homolog 3).

Paper Reference	MT Binding/MT Association	Region/Domain Indicated in MT Binding or MT Association
1. Jason Walenta et al, The Journal of Cell Biology, Volume 152, 2001	•MT association shown by immunoflourescence, labeling and imaging. Immunofluorescence microscopy localization data showed that Hook 3 accumulated close to the MTOC •Immunoprecipitation showed that endogenous hook proteins do not heterodimerise and showed MT association •MT binding shown by co-localization in a spin down assay (in vitro purified system)	•Comparison by homology showed highest degree of sequence identity in the acidic amino terminal globular domain • An extended central coiled coil motif is conserved •An amino terminal domain of 164 aa was sufficient to bind to microtubules
2. Helmut Kramer and Meridee Phistry, the Genetics Society of America, 1999	•Sequence homology searches of EST database using Blast sequence analysis of different hook alleles revealed mutations •Most mutations resulted in truncated hook protein with C-terminal deletions of various lengths •Western Analysis confirmed that an in frame deletion in the hk14 allele, resulted in the loss of 219-318 aa region of the protein. An important function of hook is associated with this deleted fragment.	•Additional homologs Hook1 and Hook2 were identified from the EST database. A 125 aa amino terminal domain and a centrally located coiled coil domain exhibited notably high conservation with 49% identity to D. melanogaster hook protein. •The hook protein has 3 domains: amino terminal domain, the C - terminal domain and a central coiled coils region of 200 aa, these are conserved in 2 drosophila and 2 human hook proteins

Table:2

III.B. MAP-DB

Four screens/web-pages were created to disseminate the microtubule binding or associated protein information through a web site. This web site is currently published using an installation of the Apache Tomcat web server on Shelob. Shelob is an evolving platform provided by the Department of Computer and Information Science at IUPUI, Indianapolis. The Shelob server was built to provide users stable and secure access to emerging web technologies. Its users are provided with an independent, configurable and self-contained server environment. It has the Tomcat 4.1 installed; complying with Servlet spec 2.3 and JSP spec 1.2. Each user has web access to their own Tomcat Administration and Tomcat Manager. User's may start and stop their instance using StartTomcat.sh and StopTomcat.sh [20]. Users also have access to their own logs and debugging information. The dynamics of the web pages and the flow of information through the web pages is depicted using the example of the microtubule binding protein, ‘Ensconsin’.

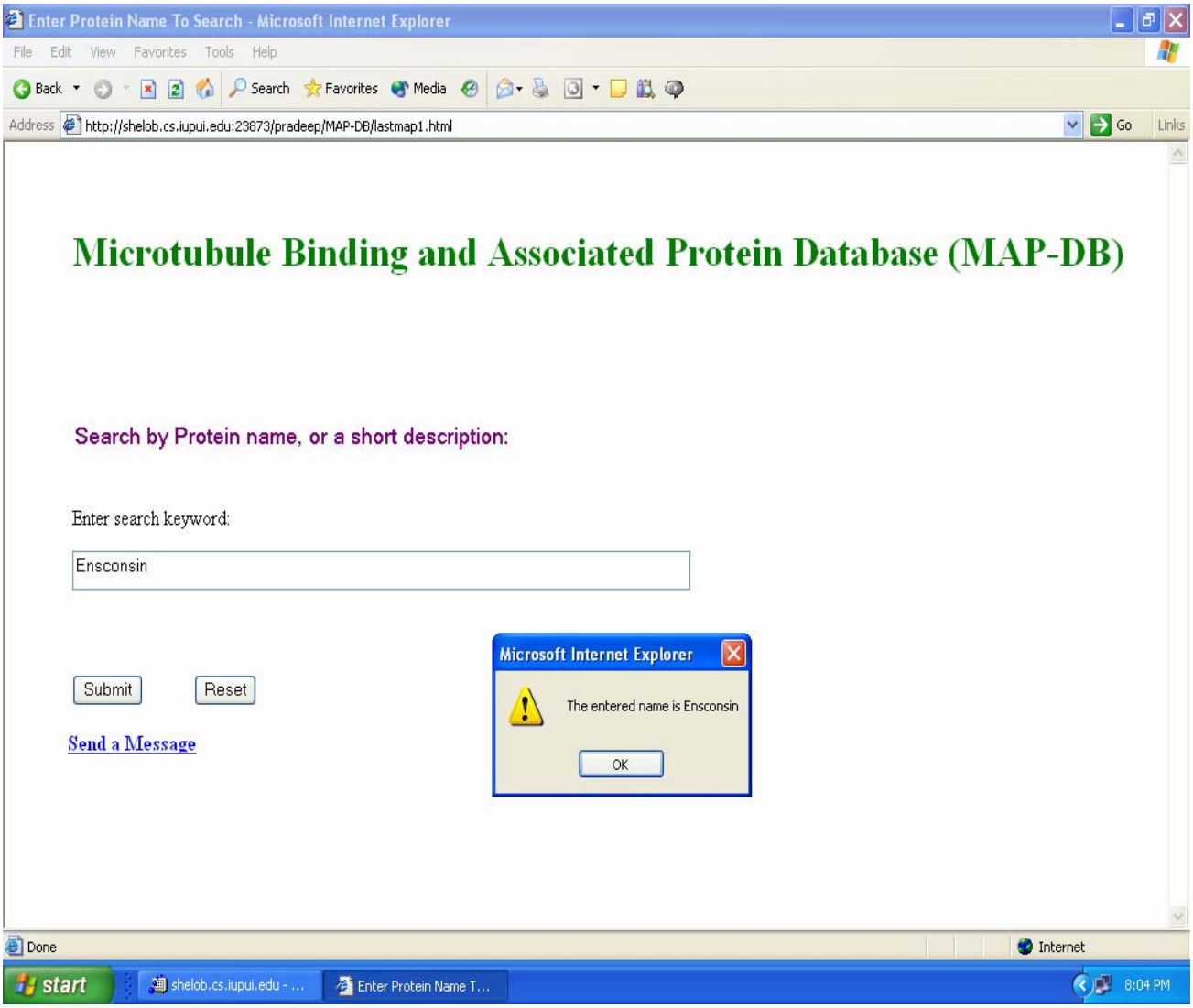


Figure 18: Screen to search for a particular protein

The screen in Figure 18, requests the user to search for a protein by its name or synonym. The user needs to input his/her search keyword in a text box and hit the submit button. A pop-up alert comes up to verify the name of the protein entered by the user. He/she also has the option to reset/clear the entry if necessary. At the bottom of the page a hyperlink with an email address is provided to enable the user to send his/her comments via email. When this link is clicked it opens the email editor on the user's system, using which comments can be emailed. The first page is a html file that calls the next screen, a JSP page.

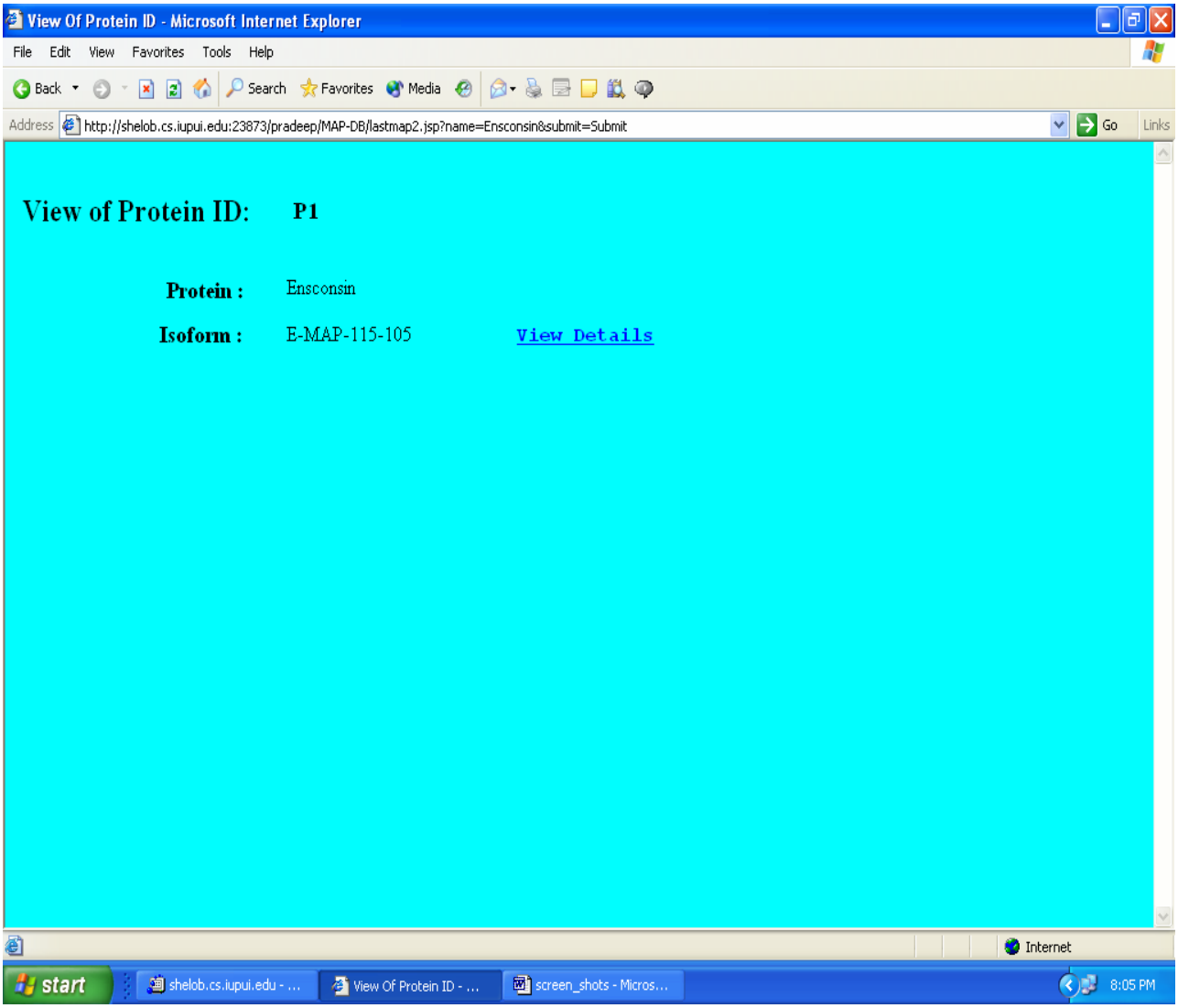


Figure 19: Screen that lists isoforms of the protein

As shown in Figure 19, for the protein searched in the first page, the names of its isoforms are displayed on the second page. Beside each isoform name a 'View Details' link is displayed. By clicking this link for a particular isoform, the user navigates to the next page (the third screen). This screen displays detailed information for that specific isoform.

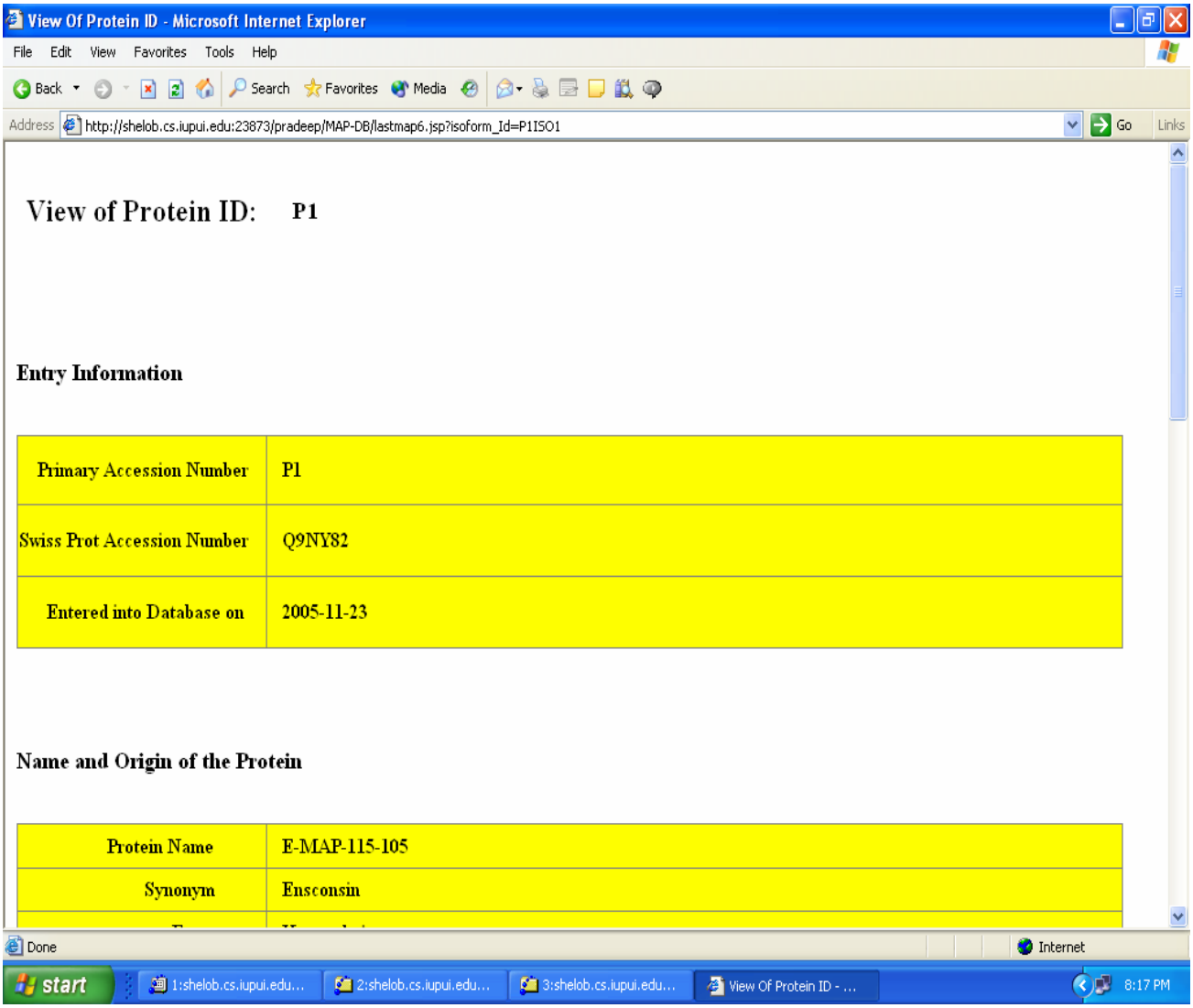


Figure 20: Screen that displays details of the protein

Information is displayed on the screen as shown in Figure 20, under the following headings: 'Entry Information', 'Name and Origin of the Protein', 'Sequence Information' and finally 'Annotation'. Under 'Entry Information', the primary accession number (generated by the curator of the database), Swiss Prot accession number, and the date the information for this specific isoform was entered into database is listed. Below the 'Name and Origin of the Protein' heading, the isoform name, any synonym/synonyms that it is known by, the name and taxonomical nomenclature of the organism from which it is obtained and the isoform's domain information is listed.

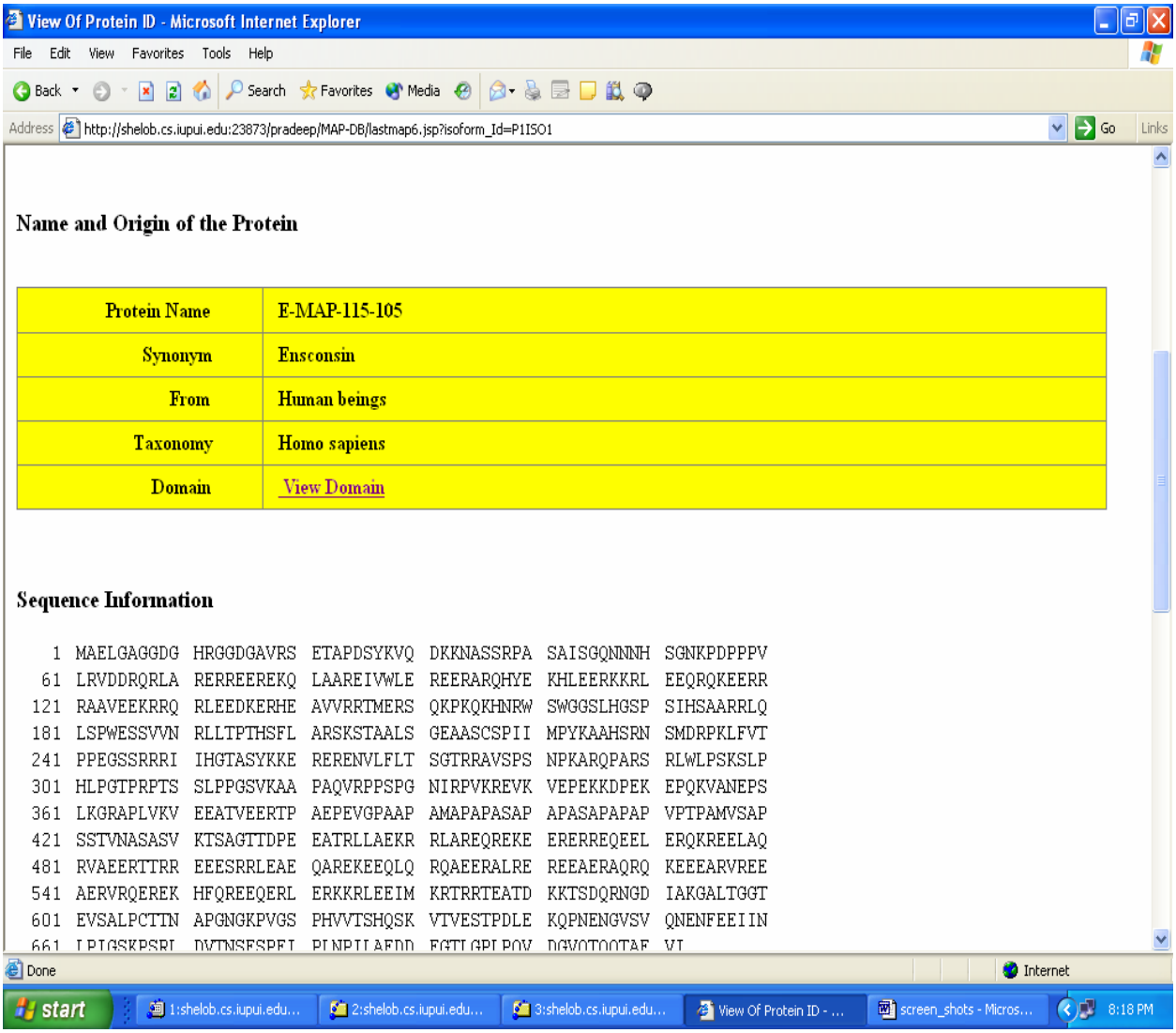


Figure 21: Sequence Information

As shown in figure 21, against the row titled ‘domain’, a link is provided that facilitates the user to view details of the domain on a fourth screen. The amino acid sequence of the isoform is displayed subsequently. Each row displays 60 amino acids (shown in sets of 10 with a space separating each set). At the beginning the number of the first amino acid in that row is printed (so, the first column in this format will have entries: 1, 61, 121, 181 and so on). This allows the user to maintain a count of the amino acids in each row and the whole sequence as well. If the user wishes to analyze amino acids in specific parts of the sequence, this format facilitates the user to select, cut and paste those

parts of the sequence that are of interest, as he/she is aware of the numbering of the amino acids within the sequence just by looking at it in this format.

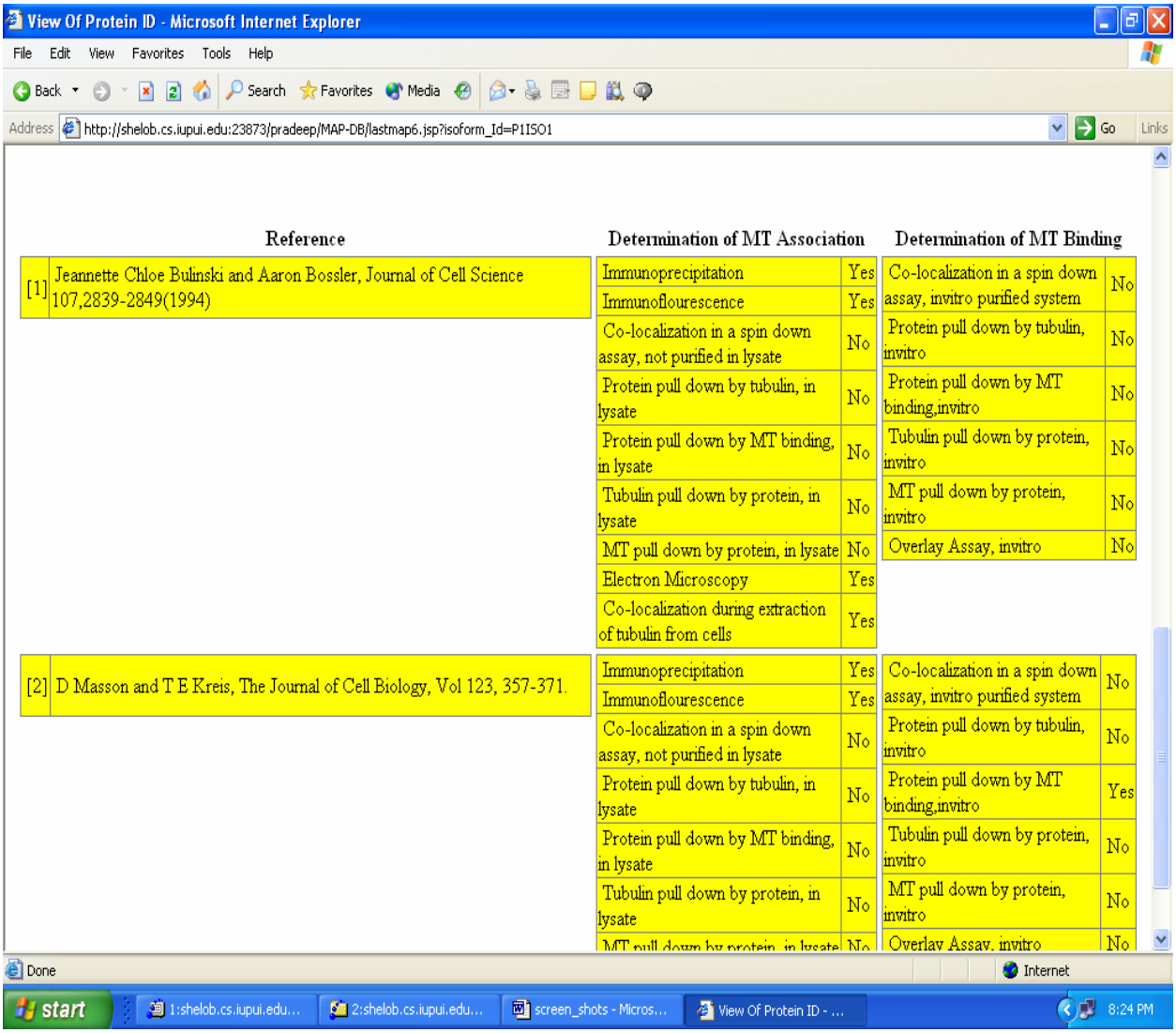


Figure 22: References and Annotation

Finally, annotation information is displayed. Every reference that was researched and annotated for the isoform is listed. Against each reference in separate adjacent tables, methods used in the determination of MT binding or MT association is listed. Against each method it is indicated whether or not, the method was employed by the reference to determine MT binding or MT association of the protein. This is shown in Figure 22.

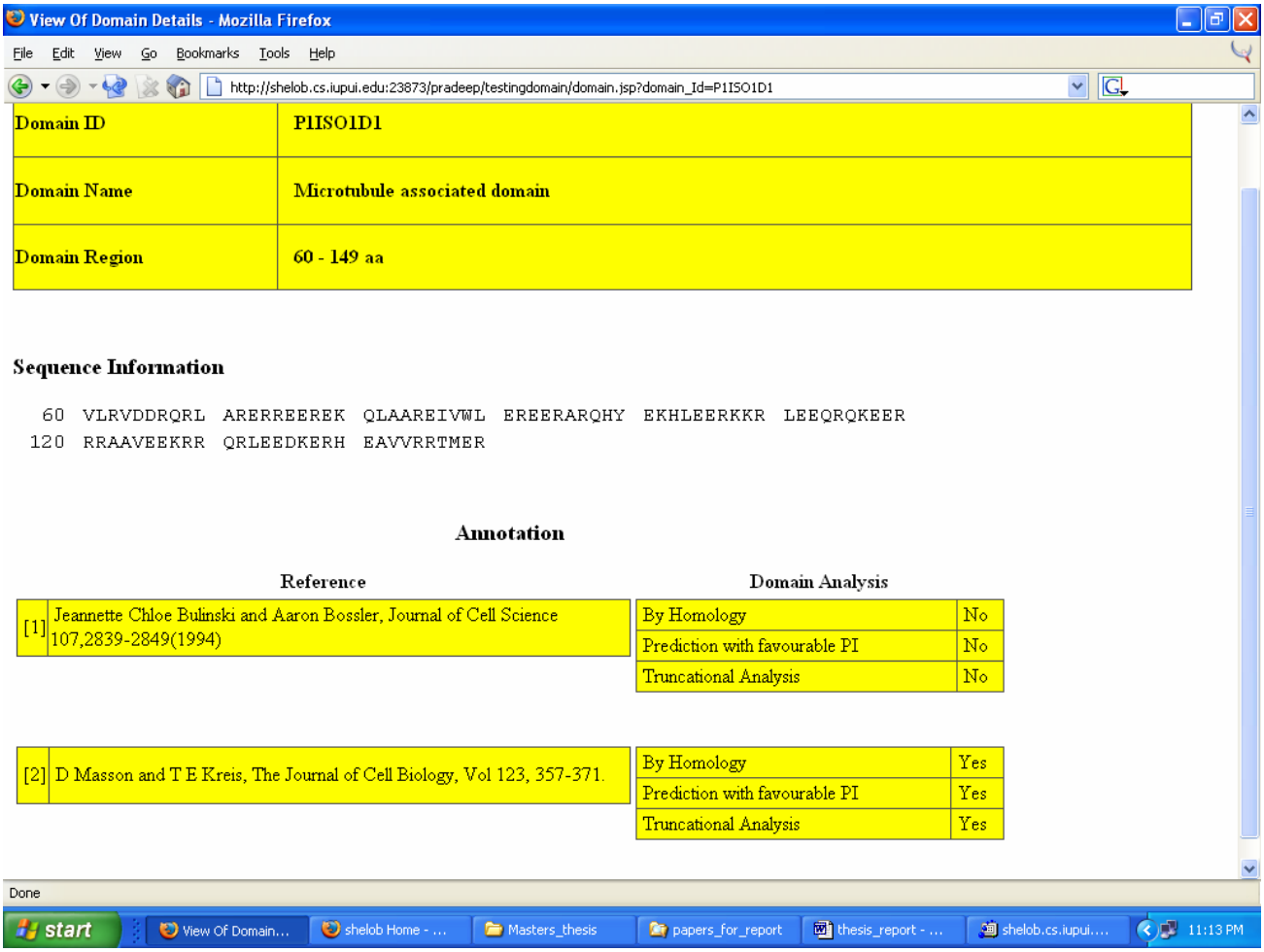


Figure 23: Domain details

As shown in Figure 23, the screen displays the details of the domain selected by the user. It lists the name of the domain, range of amino acids in the domain, the domain sequence and annotation information of the domain. The sequence information and the annotation are illustrated in a manner similar to that described in Figure 21 and 22.

IV. Discussion

The microtubule infrastructure is responsible for many aspects of cellular life such as motility, division, chromosomal separation, morphology, polarity, structure of flagellum and cilia, intracellular organization and transport. The fundamental unit in all of these different structures and functions is tubulin polymerized into microtubules. In order to mold and alter the MTs for all of these roles, the cell relies upon a host of MT associated proteins. Therefore, fundamental to understanding biological form and function, is the identification and characterization of the suite of Microtubule Associated Proteins (MAPs) involved in each MT system (such as basal bodies, centrioles, a cytosolic MT network, flagellum, kinetochores, and mitotic spindles). MAPs are vital networks of proteins that direct cellular behavior through their ability to bind microtubules (MTs) in a spatial- and temporal-specific manner and present a molecular portrait of the different MT systems within the cell.

IV.A. What we've achieved - Analysis and Annotation of MAPs and the MAP-DB

The primary goal of this research was to annotate MT binding and MT associated proteins from eukaryotes, collect information on known MAPs within these organisms and develop a database of annotated MAPs and to disseminate this information as a web resource for the scientific community. Genomic and proteomic information have led to a rapid increase in the identification and biochemical characterization of MAPs but there is no central database correlating this information within each MT system. Ours a first step in creating a DB on proteins of MT systems. The database was developed, four MAPs were annotated and the researched information now resides in the database, which we call MAP-DB. Eventually when more MAPs are added to the database, a researcher working on any MT system within the database will be able to find useful information

regardless of the organism that they are studying. This DB aims to eventually include MAPs within the Apicomplexan (Plasmodium, Toxoplasma, *etc.*) and Trypanosomatid parasites. These parasites cause deadly diseases like Malaria and Sleeping Sickness. Our annotation process will be tested when the MAPs in these parasitic systems are analyzed and this will prepare the DB for addition of other novel MT systems, such as those contained with plants. The DB will also allow us to screen out MT-binding proteins from these parasites that they share in common with humans. This will leave us with a subset of parasite specific MT-binding proteins which are potential specific drug targets. A bioinformatics approach was used to analyze and annotate the MAPs. The focus was on using bioinformatics tools and resources to identify MAPs involved in creating and maintaining unique MT structures, on characterizing how these proteins alter and stabilize the MT structures and provide the opportunity for selective molecular intervention. A protocol of analysis and annotation was devised. Bioinformatics tools such as Psi-blast, PredictProtein, Clustal W were used and literature research using PubMed was done in an attempt to identify and characterize the microtubule binding or microtubule associated domain of the MAPs and to disseminate this information without ambiguity through the MAP-DB database. There were advantages and limitations in using these tools. Therefore, an initial analysis was charted and as the analysis progressed, improvements were made and a modified protocol evolved.

IV.B. Limitations Run into in this Study

Psi-blast provided an enormous advantage over normal blast in the detection of distantly related sequences. It provided an automated, easy-to-use version of a "profile" search, which is a sensitive way to look for sequence homologues. Psi-blast aligns significantly more residues correctly whereas Clustal W did not perform as well. This was inferred in the analysis of multiple sequence alignment of homologs (picked from Psi-blast results) with the query sequence in the case of all the four MAPs analysed. The Clustal W analysis of Ensconsin and its homologs are discussed in the result section.

By the secondary structure prediction analysis of Ensconsin clear boundaries could be drawn and conserved regions or regions of apparent order could be flagged. However, secondary structure prediction did not work well for Hook (homolog 3). This implies that there are limitations in identifying conserved regions in protein structure by secondary structure analysis. This adds a cautionary note for future analysis. Computational tools are only getting better everyday, but they should not be the only methods of analysis employed. This is reiterated by the fact that in our research literature analysis helped identify the (60-149 aa) region in Ensconsin and the (1-164 aa) region as MT binding. It is also important to note how the determination of MT binding or MT association of the MAP was done at the molecular level. The MAP-DB provides this information concisely against each reference for each protein entry.

IV.C. Future Directions

There are plans to add information on whether controls were used or not in determining MT binding or MT association. Information disseminated in this manner will encourage parallel research on MAPs at the molecular level both at home and elsewhere. The point is that by coupling molecular biology and biochemical diagnostic tools with computational biology this research should evolve the development of a technology that can reasonably and accurately identify MT binding and MT associated domains in MAPs. There is scope to expand features and functionalities within the DB itself. Presently it gives details about the analyzed protein, its isoform/isoforms and the domain/domains flagged as MT binding or associated. In future, residues included in MT binding could be identified. Structures of MT binding domains could be displayed. Sophisticated animation could be used in these depictions to show the MT binding or association and the user could be provided with the option of blasting his/her query sequence against protein sequences in the database. Communication with Dr. Guenther's laboratory is encouraged in MAP-DB by the provision of a link to send comments by email. This will provide the means to communicate with

other research teams in related fields of bioinformatics and biomedical research within the scientific community. The primary challenge of this effort is to keep the information in the MAP-DB database authentic current, and updated. A future objective is to perform a primary sequence search of these genomes with each unique MAP family defined by annotation, in order to provide additional annotation and constantly evolve as more annotated data becomes available.

V. Reference

1. Mechanisms and Molecules of the Mitotic Spindle, Sharat Gadde and Rebecca Heald, Current Biology, Vol. 14, R797–R805, September 21, 2004.

2.Chromosome-Microtubule Interactions During Mitosis, J. Richard McIntosh, Ekaterina L. Grishchuk, and Robert R. West, Annual Review of Cell and Developmental Biology Vol. 18: 193-219 (Volume publication date November 2002) .

3. Accessory protein regulation of microtubule dynamics throughout the cell cycle, Lynne Cassimeris, Current Opinion in Cell Biology 1999, 11:134–141

4. Dynamics of the microtubule oscillator: role of nucleotides and tubulin- MAP interactions, E M Mandelkow, G Lange, A Jagla, U Spann and E Mandelkow **,** Max-Planck-Unit for Structural Molecular Biology, Hamburg, FRG.

5. Introduction to Protein Structure, Carl Branden and John Tooze, Garland Publishing Inc, 1991.

6.http://www.microsoft.com/office/frontpage/prodinfo/overview.mspx - Microsoft Frontpage Overview.

7. http://www.coreservlets.com/Apache-Tomcat-Tutorial/#Java-Home – Online Tutorial on Servlets.

8. http://jakarta.apache.org/tomcat/ -The Apache Tomcat web-site.

9. 'CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice.', Julie D. Thompson, Desmond G. Higgins and Toby J. Gibson, http://thr.cit.nih.gov/clustalw/clustalw.html

10.'Neurotoxic Calcium Transfer from Endoplasmic Reticulum to Mitochondria Is Regulated by Cyclin-Dependent Kinase 5-Dependent Phosphorylation of Tau', Frédéric Darios, Marie-Paule Muriel, Myriam Escobar Khondiker, Alexis Brice, and Merle Ruberg, The Journal of Neuroscience, April 20, 2005, 25(16):4159-4168

11. http://www.cytochemistry.net/Cell-biology/cilia.htm

12. ‘Tubulin and microtubule structure’ Kenneth H Downing and Eva Nogales, Current Opinion in Cell Biology 1998, 10:16-22

13. http://www.cytochemistry.net/Cell-biology/microtubule_structure.htm

14. ‘Structural Insights into Microtubule Function’, Eva Nogales**,** Annual Review of Biochemistry, Vol. 69: 277-302 (Volume publication date July 2000)

15. Vanier M T et al, Cell Motil Cytoskeleton. 2003 Aug;55(4):221-31

16. Cristiana Mollinari et al, The Journal of Cell Biology, Volume 157, June 24, 2002 1175-1186

17. Jason H. Walenta et al, The Journal of Cell Biology, Volume 152, Number 5, March 5, 2001 923–934

18. Isabelle Loı¨odice et al, Molecular Biology of the Cell Vol. 16, 1756–1768, April 2005

19. http://gogarten.uconn.edu/mcb221/class24.html

20. http://shelob.cs.iupui.edu/

21. Claire E Walczak, Current Opinion in Cell Biology 2000, 12:52–56

VI. Appendix

Resume

Narmada Shenoy

Ph: 602-570-8369(C)
Email: narmada_shenoy@hotmail.com

I have a good understanding of computational analysis of biological data, bioinformatics principles, algorithms and software for sequence alignment, similarity search of sequence databases, and functional inference. My technical knowledge includes Perl, C++, Java Server Pages, Microsoft Access, Microsoft FrontPage and Oracle.

Project Experience at Graduate Level

Application of the Genetic Algorithm in Protein Folding Simulation: Developed a Genetic Algorithm in C++, to simulate protein folding based primarily on H-H interactions. This focused on different parameters of the algorithm like bit representation scheme, mutation, crossover and population size and their impact on protein folding. A simple two-dimensional lattice was used for the folding. The project also investigated the potential usefulness of the genetic algorithm for finding the functional conformation of proteins.

The Chart Smart Database: Designed a patient charting database for use by health care professionals with privileges to view patient data. Target audience: nurses, physicians, pathologists, microbiologists, researchers, and other health professionals who need access to patient chart information. The database was created using front end Microsoft Access and back end Oracle. The database included input screens for recording data, a system for storing and retrieving data based on queries and reports for displaying these requests. The GUI was designed in a user friendly manner with focus on accurate data retrieval.

Microtubule (MT) Binding and Associated Protein Application: As part of the Masters Project I designed and developed a comprehensive database containing MT-associated proteins of the Apicomplexa and Trypanosoma genomes. The Research focus was to test and expand the annotation methods of proteins since these single cell organisms contain several unique MT based structures. The database is web enabled and uses Oracle as the back-end and MS FrontPage as the front end application. The web-site was implemented using Apache Web Server with JSDK and JDBC technology. The purpose of the web-site is to disseminate information that is well researched, documented and well presented to the scientific community. I used the following bioinformatics tools and web sites in this project: Swiss Prot, Genbank, psi-blast (NCBI), protein secondary structure prediction tools, and Clustal-W.

Presented a poster titled ‘Pilot Study for Proteomic Analysis of Toxoplasma Microtubule-Associated Proteins’ at the Proteomics Symposium on Oct 15th ’04 at Indiana University, Bloomington, IN.

Presented a poster titled 'The Microtubule Molecular Portrait of Eukaryotic Parasites Identifies Novel Drug Discovery Targets' at the 2005 Solutions Conference hosted by the IUPUI Solutions Center on March 15th '05 at Indiana University (IUPUI), Indianapolis IN.

Summary of Course Work

Informatics Management, Introduction to Bioinformatics, Introduction to Informatics
Database Systems, Statistical Methods, Basic Human Genetics, Masters Project, Independant Study, Seminar in Bioinformatics, Informatics Project Management.

Work Experience

June 2005 to Present
Intramural Research Training Award (IRTA) Student Fellow at the Laboratory of Biological Modeling (LBM/NIDDK) NIH, Bethesda MD.

- Circulating hormones in the gut influence energy homeostasis, the project goal is to understand the network of proteins involved in gut-hormone appetite regulatory systems. These systems are potential targets for the design of antiobesity drugs.
- Protein-protein interactions are studied and mapped using bioinformatics and mathematical tools like STRING and Pajek.

June 1996 – May 1998
Analyst, Quality Control Department
Overseas Pharma, Bangalore, India

- Conducted microbial assays of compounds such as B12, Ca panthotenate, and carried out microbial counts as part of quality compliance processes laid down by federal regulatory standards.
- Analyzed strength of chemical components in drugs using gravimetric and volumetric techniques to verify composition in finished products.
- Performed identification of substances in raw materials and finished products using physical and chemical methods such as potentiometric titration, thin layered chromatography, and assay by ultraviolet and visible absorption spectrophotometry.
- Documented analytical and test data in accordance with Federal validation standards.

Education

Master of Science in Bioinformatics, Indiana University, GPA 3.6 I graduate in August 2005.
Master of Science in Botony, Bangalore University, India, May 1996.
Bachelor of Science in Microbiology, Bangalore University, India, March 1994.

Other:

I am a Permanent Resident of the United States.